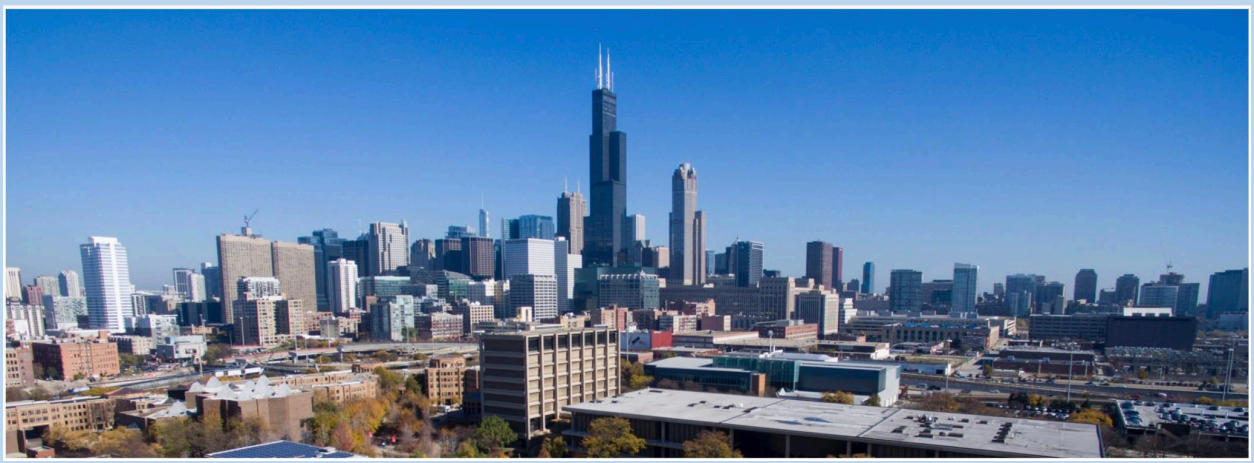


Spatiotemporal Big Data Analysis of Opioid Epidemic in Illinois

Team Name:
UI Health

Written/Submitted By:
Arash Jalali, MPH, MSHI
jalali@uic.edu

Sean Huang, MD
sshuang@uic.edu

Challenge Category II

UI Health | UIC

Abstract

In 2019, 855 people died in Chicago from opioid overdoses, more than homicides or all traffic crashes that year, and about 1.5 times higher than the rest of Illinois. This is a dramatic increase over the previous 10 years,[1] as the overall opioid-related overdose death rate increased by 10.1%. Likewise, the Chicago Fire Department emergency medical services team responded to an average of 29 responses per day in 2019 (or an 25.4% increase over the same time period). This suggests more awareness about calling for help during an overdose[1].

While not all overdoses are fatal, identifying these patients is important. Knowing patient location and demographics can help with allocating resources for treatment of substance use. We do have public health information about which segment of the Chicago population is most affected by opioid-related overdose deaths. In 2018, death rates in the city were highest among men, those aged 45-64, and non-Hispanic African-Americans[1]. The latter is in contrast to the United States as a whole where the highest drug overdose death rate is among non-Hispanic white patients. These deaths often involved a combination of opioids with other illicit drugs, cocaine in particular.

Many of those affected by opioid-related overdose death rates experience high economic hardship. This includes education, income levels, and crowded housing. Unemployment is highly correlated as well. On the other hand, work-related injuries also play a large role especially when it comes to prescription opioid pain relievers. According to John Howard and the National Institute for Occupation Safety and Health, the potential for addiction may be preceded by injuries that happen in the workplace, with the consequences affecting both an individual's working life as well as their home life.[2] Exposure to opioid powders can also create an unsafe hazardous environment for EMTs and healthcare workers.

Part of understanding the pervasiveness of the opiate crisis is in the city's geography. Availability of heroin in the Chicago area remains high and has increased over the past 10 years due to large supply and low price[3]. Most of these drugs were obtained through illicit means and deaths usually were secondary to illicit drug use (heroin, fentanyl) rather than prescription pain relievers. Likewise, Chicago serves as the primary distribution hub for opioids and other illegal drugs in the surrounding area and states. Many of the drugs moving into the region are shipped by way of trucks, buses, delivery services, and personal vehicles after being smuggled into the country[3]. Chicago is one of the country's largest trucking centers with easy access to numerous trucking depots and many large interstates. The west side of the city is probably the region's most significant market, in part due to its accessibility by Interstate 290[3]. The infrastructure allows opiates to disperse into neighboring rural areas.

Description of the Solution

As a result, we would like to explore spatiotemporal distribution of emergency 911 calls and ambulance dispatches related to drug overdoses. Chicago EMS data is obtained and stored in a Data Lake while scripts are executed over Azure Cloud services. From this data, opiates cases are identified and geospatial information is extracted, then analyzed in ArcGIS enterprise. We will then enrich this data with information based on demographic and census data of the surrounding neighborhoods. Machine learning models will be run to better understand features that can predict opioid use. This process can be greatly augmented using synthetic data generated by Synthea. Using data obtained from these calls, we can get a better sense of which communities are affected, allowing us to help allocate resources to those patients who are at the greatest in need.

Methods

Chicago EMS data was obtained from the Illinois Department of Public Health with a database structure using the National Emergency Medical Services Information System (NEMSIS) standard. This information is then transferred using an encrypted campus express route to an Azure Data Lake. Next, we ran Azure Synapse pipelines to transform the NEMSIS and Synthea generated data into an Azure SQL data warehouse. All our interactions occurred within a HIPAA-approved cloud computing environment. All the machine learning services are executed over private links using an Azure private endpoint networking infrastructure.

From this data, opioid cases are identified on the SQL server. Opiate cases are identified as any patient where the EMS 'Provider's Primary Impression' is listed as 'Opioid related disorders' or 'Opioid use, unspecified' or the 'Primary Symptom' is listed as 'Opioid use, unspecified.' Likewise, we include patients where either category matched a F11, T40, or Z79.891 ICD-10 code.

After the data in NEMSIS is cleaned, it is fed into ArcGIS Enterprise to help with geospatial data management, data visualization, analytics, and geospatial forecasting of opiates and overdoses in different areas. Utilizing ArcGIS Pro, space time cubes were generated from SQL data warehouse tables for all overdoses and opioid cases. A space time pattern manning toolset was utilized to create predictive models using time series forecasting tools. We deployed curve fit forecast models which forecast future values of our generated space time cubes using curve fitting. In addition, we deployed an exponential smoothing forecast model which predicts values of each location of a space-time cube using the Holt-Winters exponential smoothing method by decomposing the time series at each location cube into seasonal and trend components[4].

The data will be stored in an Azure Data Lake. Rather than using a data warehouse architecture, a Data Lake allows us to implement data warehousing functionality over open file formats. This grants us the ability to separate computation and storage from one another, allowing for more efficiency.

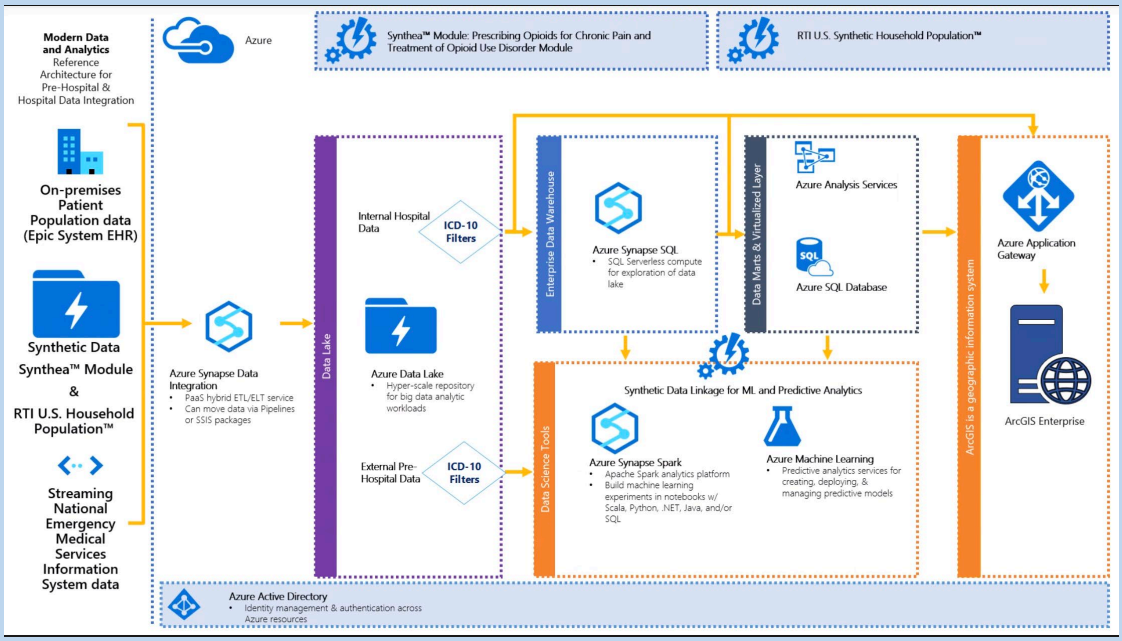


Figure description: The data analytics architecture that we used is shown here. After storing our data in the Azure Data Lake, we are then ingesting the data into Azure Synapse SQL Server, then using Azure pipelines to clean and filter the data inside the Azure SQL data warehouse. From there, we can move the data into ArcGIS Enterprise for geospatial analysis.

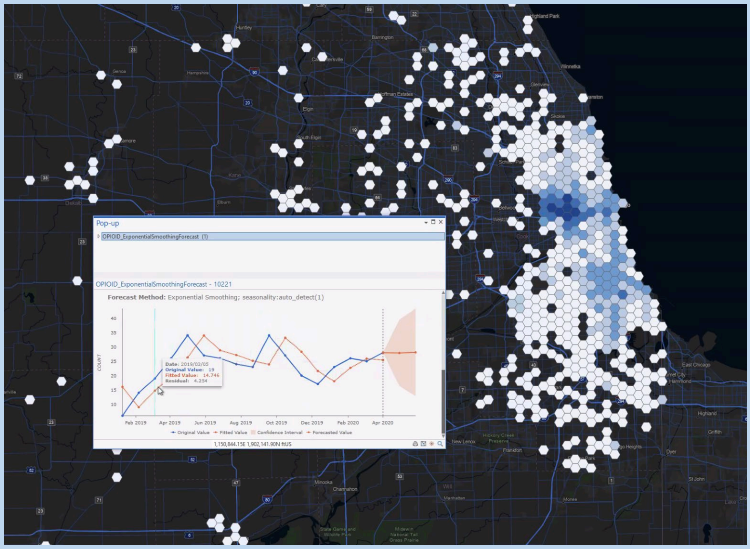
With geospatial information obtained from our data sources, the data will be uploaded into ESRI, a demographic data store that allows for geocoding and automated enrichment. Its data sources include USA 2020 demographic data, USA 2010 Census Demographic Data, USA 2014/2018 American Community Survey (ACS) Demographic Data, USA 2020 Consumer Expenditure data, and USA 2020 Tapestry Segmentation Data. Its geography information is updated to 2020/2021. Using ESRI, we will create a 5 minute walk time around each home address and calculate information.

We will enter these variables into Azure machine learning services to run predictions. Many AutoML experiments will be created in our University of Illinois Azure portal. Using a target of opioid activations by EMS and matching it with all overdose cases, we ran machine learning models using Azure Machine Learning. Multiple experiments can be run. We will use classification machine learning models. We will examine the iteration charts, evaluate their models and note accuracy, prior to deploying our model.

A similar process can be run using synthetic data generated by Synthea. Implementing this data, we find patients with addresses, zip codes, but more important for our project, latitude/longitude or x,y coordinate data. These coordinates can be matched and enriched with ESRI data in a similar manner to how we matched NEMSIS data. Afterwards, we can add social determinants of health data to our synthetic data and run similar models. This establishes a new application of Synthea generated data as it allows us to generate public health related prediction models for clinical outcomes. Here we can establish a likelihood for opioid use when combined with our real-time NEMSIS data to see which patients may need to be targeted for increased opiate mitigation resources.

Results

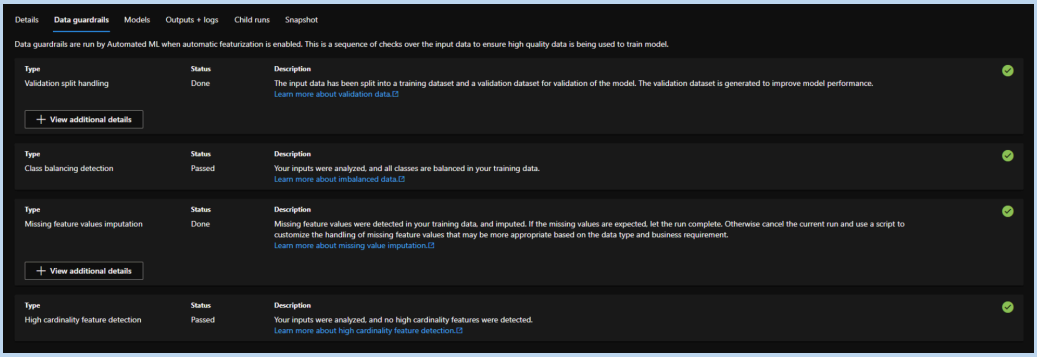
The NEMSIS data was fed into geolocation services in ArcGIS Enterprise to help forecast and visualize opiates and overdoses. We can see similar results with all overdoses. In the following analysis, we see a peak in overdose cases in the summer months. The map shows also higher incidence of overdoses in the west side of the city of Chicago. In the second map on the next page, we see strong forecasting of opioid cases using our deployed forecasting tools. Here we see that the blue line (which indicates actual cases) and the orange line (which indicates forecasting based on geospatial data) match well. We can use this method to consider where to deploy our resources.



Here are examples of curve data plots for overdose and opioid data. In the first picture, we see a forecasted value that is close to 19 overdoses cases in this particular area. In the second picture, we demonstrate using machine learning to predict another 21-22 cases in the next month. Equations can be obtained as well for these best fit lines.

OBJECTID *	Shape *	Location ID	Forecast for COUNT in 2020-04-28 23:28:00	Forecast for COUNT in 2020-05-26 23:28:00	Forecast Root Mean Square Error	Validation Root Mean Square Error (Validation Steps: 2)	Forecast Method	Forecast Equation
1	Polygon	53	-0.029412	-0.039216	0.23034	0.024986	linear	Xt = a*t+b; a=-0.009804, b=0.137255
2	Polygon	303	-0.077356	-0.137483	0.232958	0.066017	exponential	Xt = k + a*exp(b*t); k=0.084258, a=-0.000747, b=0.316302
3	Polygon	642	0.014706	0.029412	0.221541	0.004806	parabolic	Xt = a*t^2+b*t+c; a=0.001548, b=-0.039474, c=0.238390
4	Polygon	643	-0.117647	-0.205882	0.303763	0.043743	parabolic	Xt = a*t^2+b*t+c; a=-0.004902, b=0.083333, c=-0.117647
5	Polygon	644	-0.044118	-0.088235	0.227489	0.034787	parabolic	Xt = a*t^2+b*t+c; a=-0.002580, b=0.046182, c=-0.083591
6	Polygon	645	0.224392	0.224568	0.306001	0.621514	gompertz	Xt = k+a*exp(-b*exp(-c*t)); k=0.000000, a=0.224779, b=52.858909, c=0.6076...
7	Polygon	962	-0.044118	-0.088235	0.227489	0.034787	parabolic	Xt = a*t^2+b*t+c; a=-0.002580, b=0.046182, c=-0.083591
8	Polygon	1184	0.014706	-0.014706	0.22882	0.153306	parabolic	Xt = a*t^2+b*t+c; a=-0.001935, b=0.038313, c=-0.077399
9	Polygon	1193	-0.007353	-0.014706	0.23252	0.006942	linear	Xt = a*t+b; a=-0.007353, b=0.117647
10	Polygon	1208	0.213235	0.230392	0.21977	0.707107	linear	Xt = a*t+b; a=0.017157, b=-0.078431
11	Polygon	1210	1.235294	1.45098	0.548725	1.035503	parabolic	Xt = a*t^2+b*t+c; a=0.008256, b=-0.073271, c=0.094943
12	Polygon	1211	0.491678	0.511884	0.271131	0.525169	gompertz	Xt = k+a*exp(-b*exp(-c*t)); k=0.000010, a=0.557301, b=91.371438, c=0.3877...
13	Polygon	1389	-0.007353	-0.014706	0.23252	0.006942	linear	Xt = a*t+b; a=-0.007353, b=0.117647
14	Polygon	1441	0.191176	0.205882	0.223993	0.279446	linear	Xt = a*t+b; a=0.014706, b=-0.058824
15	Polygon	1458	-0.176471	-0.29902	0.354186	0.088862	parabolic	Xt = a*t^2+b*t+c; a=-0.006579, b=0.107714, c=-0.106295
16	Polygon	1459	-0.176471	-0.284314	0.45257	0.138673	parabolic	Xt = a*t^2+b*t+c; a=-0.005934, b=0.099845, c=-0.158927
17	Polygon	1604	0.194342	0.194347	0.30925	0.256445	gompertz	Xt = k+a*exp(-b*exp(-c*t)); k=0.000000, a=0.194351, b=108.978494, c=0.859...
18	Polygon	1637	0.588235	0.696078	0.359418	0.551234	parabolic	Xt = a*t^2+b*t+c; a=0.004902, b=-0.063725, c=0.254902
19	Polygon	1660	-0.077356	-0.137483	0.232958	0.066017	exponential	Xt = k + a*exp(b*t); k=0.084258, a=-0.000747, b=0.316302
20	Polygon	1661	0.147059	0.156863	0.23034	0.218635	linear	Xt = a*t+b; a=0.009804, b=-0.019608

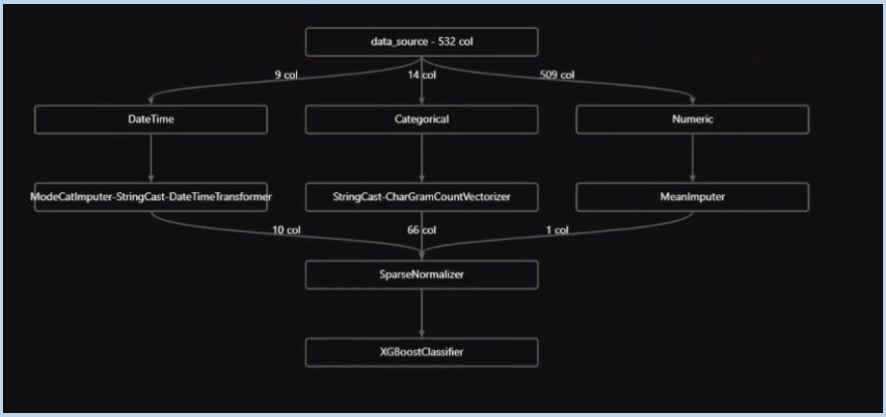
The NEMSIS data was matched with enriched data from ESRI based on location. The variables and columns were cleaned and filtered. Some variables had to be removed as they led to overfitting. Other variables were removed to imbalanced data. As noted here, the data guardrails have all been cleared and passed.



Afterwards, the experiments were run, with many models produced as shown here, along with how the data was transformed.

Algorithm name	Explained	Accuracy ↓	Sampling
VotingEnsemble	View explanation	0.84012	100.00 %
StackEnsemble		0.83637	100.00 %
SparseNormalizer, XGBoostClassifier		0.83388	100.00 %
SparseNormalizer, XGBoostClassifier		0.83388	100.00 %
SparseNormalizer, XGBoostClassifier		0.83388	100.00 %
SparseNormalizer, XGBoostClassifier		0.83341	100.00 %
SparseNormalizer, XGBoostClassifier		0.83326	100.00 %
SparseNormalizer, XGBoostClassifier		0.83326	100.00 %
SparseNormalizer, XGBoostClassifier		0.83326	100.00 %
MaxAbsScaler, LightGBM		0.83232	100.00 %
SparseNormalizer, XGBoostClassifier		0.83216	100.00 %

Algorithm name	Explained	Accuracy ↓	Sampling
MaxAbsScaler, LogisticRegression		0.82468	100.00 %
MaxAbsScaler, LightGBM		0.82405	100.00 %
MaxAbsScaler, LightGBM		0.82358	100.00 %
SparseNormalizer, XGBoostClassifier		0.82296	100.00 %
StandardScalerWrapper, LightGBM		0.82171	100.00 %
SparseNormalizer, XGBoostClassifier		0.82140	100.00 %
MaxAbsScaler, LightGBM		0.82124	100.00 %
MaxAbsScaler, GradientBoosting		0.82093	100.00 %
MaxAbsScaler, GradientBoosting		0.82093	100.00 %
StandardScalerWrapper, RandomForest		0.81813	100.00 %
SparseNormalizer, LightGBM		0.81563	100.00 %



We can highlight the third algorithm, that of SparseNormalizer, XGBoost Classifier which did produce an accuracy of 0.83388, indicating that our process is a strong predictor for our target of opiate use and drug overdoses in general. This was generated after close to 50 runs with various machine learning models, each one gaining knowledge from the previous iteration. Of note, the algorithm with the highest accuracy was VotingEnsemble. which is a combination of models designed to produce the good accuracy, however it is not as easily explainable.

Considerations with Synthea

As discussed earlier, a similar machine learning method can be applied to Synthea generated synthetic health data. As shown below, Synthea has the ability to create geographic data with x,y coordinates. We used the script shown below to generate synthetic cases for the Chicago, Illinois area

```
C:\Synthea\synthea>.\run_synthea.bat -m "onc_opioids" -p 255000 Illinois Chicago_
```

```
254984 -- Elva122 Langworth352 (25 y/o F) Chicago, Illinois
254982 -- Drucilla444 Paucek755 (29 y/o F) Chicago, Illinois
254991 -- Coleen678 Sauer652 (17 y/o F) Chicago, Illinois
254988 -- Tracey100 Gottlieb798 (30 y/o M) Chicago, Illinois
254985 -- Luigi346 Schmeler639 (39 y/o M) Chicago, Illinois
254983 -- Kevin729 Hahn503 (54 y/o M) Chicago, Illinois
254989 -- Long300 Hammes673 (41 y/o M) Chicago, Illinois DECEASED
```

		CITY	STATE	COUNTY	ZIP	LAT	LON	HE
9	rchard Apt...	Chicago	Illinois	DuPage County	60018	41.881	-87.616	
10		Chicago	Illinois	DuPage County	60646	41.774	-87.806	
11	Suite 29	Chicago	Illinois	DuPage County	60616	41.904	-87.779	
12	arade	Chicago	Illinois	DuPage County	60068	41.959	-87.617	
13	Harbor	Chicago	Illinois	DuPage County	60647	41.866	-87.642	
14		Chicago	Illinois	DuPage County	60634	41.944	-87.769	
15	ay	Chicago	Illinois	DuPage County	60640	41.640	-87.578	
16	ding	Chicago	Illinois	DuPage County	60621	41.734	-87.666	
17	Unit 29	Chicago	Illinois	DuPage County	60610	41.667	-87.629	
18	m	Chicago	Illinois	DuPage County	60661	42.001	-87.689	
19		Chicago	Illinois	DuPage County	60176	41.772	-87.562	
20		Chicago	Illinois	DuPage County	60652	41.986	-87.657	
21	unction	Chicago	Illinois	DuPage County	60610	41.972	-87.785	
22	en	Chicago	Illinois	DuPage County	60617	41.715	-87.615	
23	Suite 55	Chicago	Illinois	DuPage County	60614	41.809	-87.609	
24	Meadow	Chicago	Illinois	DuPage County	60630	41.799	-87.696	
25	rade	Chicago	Illinois	DuPage County	60604	41.919	-87.672	
26	Gate	Chicago	Illinois	DuPage County	60610	41.816	-87.776	
27	r Unit 36	Chicago	Illinois	DuPage County	60617	41.842	-87.762	
28	hroughwa...	Chicago	Illinois	DuPage County	60654	42.029	-87.621	

These x, y coordinates that were generated for a synthetic population in the city of Chicago can be mapped and enriched using ArcGIS Enterprise. We can generate 5 minute walk times around each synthetic patient's home address and enrich using USA 2020 demographic data, USA 2010 Census Demographic Data, USA 2014/2018 American Community Survey (ACS) Demographic Data, USA 2020 Consumer Expenditure data, and USA 2020 Tapestry Segmentation Data. Through this process, we are able to merge Synthea generated synthetic data with important up-to-date demographic datasets. When the data is enriched, we can move them into Azure Machine Learning for similar classification techniques.

While we were able to take advantage of this amazing capability of Synthea generated synthetic data, our machine learning results were not as robust as it was with real 911 data. We suspect that this is due to how the generated data is not reflective enough of the true demographics of the Chicago area. Such can be seen when examining the racial/ethnicity data as well as by viewing the geolocation map.

A novel future improvement to Synthea can be accomplished by integrating RTI U.S. Synthetic Household population data[5] into the Synthea workflow. Our recommendation to public health informaticians is to substitute the Synthea geography data file with the RTI geography file. This will allow for more accurate representation of the patient population in the United States. In addition, we are hopeful that NEMSIS will provide zip code information as part of its 2020 public-release research dataset. Thus, we can use the same method and public health architecture to combine pre-hospital patient records with hospital patient records with the ultimate goal of improving clinical outcomes.

References

[1] Blair Turner, Wilnise Jasmin Isabel Chung, Ponni Arunkumar, Mark Kiely, Steven Aks, Nikhil Prachand, Allison Arwady. Opioid Overdose Surveillance Report—Chicago 2019. City of Chicago, March 2021.

[2] The National Institute for Occupational Safety and Health (NIOSH) (2020, April 13) *Opioids in the Workplace.* Centers for Disease Control and Prevention. https://www.cdc.gov/niosh/topics/opioids/

[3] DEA Intelligence Report. (2017). The Opioid Threat in the Chicago Field Division (Report No. DEA-CHI-DIR-023-17). DEA United States Drug Enforcement Administration. https://www.dea.gov/documents/2017/2017-06/2017-06-01/opioid-threat-chicago-field-division

[4] Xiodan Zhou (2020, July 20) Time Series Forecasting 101 – Part 4. Forecast and visualize with Exponential Smoothing. ESRI ArcGIS Blog. https://www.esri.com/arcgis-blog/products/arcgis-pro/analytics/time-series-forecasting-101-part-4-forecast-and-visualize-with-exponential-smoothing/

[5] "RTI U.S. Synthethic Household Population™" *RTI International.* https://www.rti.org/impact/rti-us-synthetic-household-population%E2%84%A2 Accessed July 2021.