Project Description

While estimating health outcomes at a neighborhood scale is important for promoting urban health, it has been a costly and time-consuming task. The *Urban Health Risk Mapping* project leverages crowdsourced data and machine learning technologies to predict the census tract-level health outcomes for ten major US cities, including Austin, Baltimore, Boston, Dallas, Washington, D.C., Houston, Los Angeles, New York City, San Antonio, and San Francisco. The machine-learning-enabled approach has an advantage over the traditional survey methods in terms of time and cost.

The project consists of four parts: (1) database development, (2) modeling and analytics, (3) visualization and web development, and (4) community engagement and application. The first two parts are associated with the actual building, training, and testing of machine learning models. The targets are the various health outcomes, namely the prevalence of common non-communicable chronic diseases such as coronary heart disease, cancer, diabetes, poor mental health, obesity, and stroke. The actual health outcomes used in training and testing the models are accessed from the CDC's 500 Cities Project. The features are created based on three data sources, namely the CDC's Social Vulnerability Index (SVI) dataset, the EPA's Smart Location Database (SLD), and the 311 service request datasets accessed from each municipality. Sixty features (i.e., predictor variables) are considered, which characterize the social environment, the physical environment, and the aspects and degrees of neighborhood disorder. A variety of machine learning algorithms are applied and compared, including Ridge Regression, Lasso Regression, Elastic Net, Support Vector Machine, Decision Tree, Random Forest, Extra Trees, and Gradient Boosting. To improve the model performance, the model hyperparameters are fine-tuned using 10-fold cross-validation. Different sets of features are also experimented with.

It is shown that the tract-level prevalence for the common non-communicable chronic diseases can be reasonably well predicted based on the publicly available datasets. Furthermore, two major findings have been yielded from this study: (1) the sociodemographic and socioeconomic variables are the strongest predictors for tract-level health outcomes; (2) the historical records of 311 service requests can be a useful complementary data source because the information distilled from the 311 data often helps improve the models' performance.

The datasets and the predictive models are published online. Users can play with the models interactively by using the web tools we developed. The web tools can help the public and city officials evaluate future scenarios and understand how changes in the neighborhood conditions can lead to changes in the health outcomes.

Data Sources

The census tract-level health data are drawn from the 500 Cities Project dataset. (https://chronicdata.cdc.gov/browse?category=500+Cities)

The built environment variables are calculated based on EPA's Smart Location Database (SLD). (https://www.epa.gov/smartgrowth/smart-location-mapping#SLD)

The socioeconomic and sociodemographic variables are extracted from CDC's Social Vulnerability Index (SVI) dataset. (https://svi.cdc.gov/data-and-tools-download.html)

The 311 data are accessed from the open data portal of each municipality.

(Austin: https://data.austintexas.gov/Utilities-and-City-Services/Austin-311-Public-Data/xwdj-i9he

Baltimore: https://data.baltimorecity.gov/City-Services/311-Customer-Service-Requests/9agw-sxsr

Boston: https://data.boston.gov/dataset/311-service-requests

Dallas: https://www.dallasopendata.com/City-Services/311-Service-Requests-October-1-2016-to-September-3/shgm-yzbp

https://www.dallasopendata.com/City-Services/311-Service-Requests-October-1-2018-to-Present-/m36q-vtbr

Washington, D.C.: https://opendata.dc.gov/datasets/311-city-service-requests-in-2019

Houston: http://www.houstontx.gov/311/

Los Angeles: https://data.lacity.org/browse?q=311&sortBy=relevance&page=2

New York City: https://nycopendata.socrata.com/Social-Services/311-Service-Requests-from-2010-to-Present/erm2-nwe9

San Antonio: https://data.sanantonio.gov/dataset/service-calls

San Francisco: https://data.sfgov.org/City-Infrastructure/311-Cases/vw6y-z8j6/data)

Table S1. Abbreviations and Descriptions of Variables.

Variable	Abbreviation	Data Source
Outcome variable		
Arthritis among adults aged ⩾ 18 years (%)	ARTHRITIS	CDC's 500 Cities Project
High blood pressure among adults aged ⩾ 18 years (%)	BPHIGH	
Cancer (excluding skin cancer) among adults aged ⩾ 18 years (%)	CANCER	
Current asthma prevalence among adults aged ⩾ 18 years (%)	CASTHMA	
Coronary heart disease among adults aged ⩾ 18 years (%)	CHD	
Chronic obstructive pulmonary disease among adults aged ⩾ 18 years (%)	COPD	
Diagnosed diabetes among adults aged ⩾ 18 years (%)	DIABETES	
High cholesterol among adults aged ⩾ 18 years who have been screened in the past 5 years (%)	HIGHCHOL	
Chronic kidney disease among adults aged ⩾ 18 years (%)	KIDNEY	
Mental health not good for ⩾ 14 days among adults aged ⩾ 18 years (%)	MHLTH	
Physical health not good for ⩾ 14 days among adults aged ⩾ 18 years (%)	PHLTH	
Stroke among adults aged ⩾ 18 years (%)	STROKE	
All teeth lost among adults aged ⩾ 65 years (%)	TEETHLOST	
Binge drinking prevalence among adults aged ⩾ 18 years (%)	BINGE	
Current smoking among adults aged ⩾ 18 years (%)	CSMOKING	
No leisure-time physical activity among adults aged ⩾ 18 years	LPA	
Obesity among adults aged ⩾ 18 years	OBESITY	
Sleeping less than 7 hours among adults aged ⩾ 18 years	SLEEP	

Note: The column names for the predicted health outcome values are made simply by prefixing a lowercase 'p' before the variable names shown above. For example, 'ARTHRITIS' becomes 'pARTHRITIS'.

Predictor variable

Percentage of persons below poverty	P_POV	CDC's SVI data
Percentage of civilian (age 16+) unemployed estimate	P_UNEMP	
Per capita income (US$)	PCI	
Percentage of persons with no high school diploma (age 25+)	P_NOHSDP	
Percentage of persons aged 65 and older	P_AGE65P	
Percentage of persons aged 17 and younger	P_AGE17M	
Percentage of civilian noninstitutionalized population with a disability	P_DISABL	
Percentage of single parent households with children under 18	P_SNGPNT	
Percentage minority (all persons except white, non-Hispanic)	P_MINRTY	
Percentage of persons (age 5+) who speak English "less than well"	P_LIMENG	
Percentage of housing in structures with 10 or more units	P_MUNIT	
Percentage of mobile homes	P_MOBILE	
Percentage of occupied housing units with more people than rooms	P_CROWD	
Percentage of households with no vehicle available	P_NOVEH	
Percentage of persons in institutionalized group quarters	P_GROUPQ	
Percentage uninsured in the total civilian noninstitutionalized population	P_UNINSUR	
Percent of population that is working aged	P_WRKAGE	EPA's Smart Location Database
Percent of one-car households	P_AO1	
Percent of two-plus-car households	P_AO2P	
Percentage of low-wage workers (earning $1250/month or less) among total workers (home location)	P_LOWWAGEr	
Percentage of low-wage workers (earning $1250/month or less) among total workers (work location)	P_LOWWAGEe	

Gross residential density (HU/acre) on unprotected land	D_HH
Gross population density (people/acre) on unprotected land	D_POP
Gross employment density (jobs/acre) on unprotected land	D_EMP
Gross activity density (employment + HUs) on unprotected land	D_HUEMP
Gross retail (5-tier) employment density (jobs/acre) on unprotected land	D_EMP_RET
Gross office (5-tier) employment density (jobs/acre) on unprotected land	D_EMP_OFF
Gross industrial (5-tier) employment density (jobs/acre) on unprotected land	D_EMP_IND
Gross service (5-tier) employment density (jobs/acre) on unprotected land	D_EMP_SVC
Gross entertainment (5-tier) employment density (jobs/acre) on unprotected land	D_EMP_ENT
Jobs per household	JOBSPERHH
5-tier employment entropy (denominator set to observed employment types in the census tract)	EMPMIX
Employment and household entropy	EMPHHMIX
Employment and household entropy (based on vehicle trip production and trip attractions including all 5 employment categories)	TRIPMIX
Trip productions and trip attractions equilibrium index	TRIPEQ
Household workers per job, by census tract	WRKSPERJOB
Household workers per job equilibrium index	HHWRKJOBEQ
Total road network density	D_RD
Network density in terms of facility miles of auto-oriented links per square mile	D_RD_AO
Network density in terms of facility miles of multi-modal links per square mile	D_RD_MM
Network density in terms of facility miles of pedestrian-oriented links per square mile	D_RD_PO
Street intersection density (auto-oriented intersections eliminated)	D_X_EXCLAO
Intersection density in terms of auto-oriented intersections per square mile	D_X_AO

Intersection density in terms of multi-modal intersections having three legs per square mile	D_X_MM3
Intersection density in terms of multi-modal intersections having four or more legs per square mile	D_X_MM4
Intersection density in terms of pedestrian-oriented intersections having three legs per square mile	D_X_PO3
Intersection density in terms of pedestrian-oriented intersections having four or more legs per square mile	D_X_PO4
Proportion of census tract employment within ¼ mile of fixed-guideway transit stop	P_EMP025
Proportion of census tract employment within ½ mile of fixed-guideway transit stop	P_EMP050
Aggregate frequency of transit service per square mile	D_TRANSIT