CITY-LEVEL ADULT STROKE PREVALENCE IN RELATION TO REMOTE SENSING DERIVED PM2.5 ADJUSTING FOR UNHEALTHY BEHAVIORS AND MEDICAL RISK FACTORS

Z. Hu[1, *]

[1] Department of Earth & Environmental Sciences, University of West Florida, Pensacola, FL, 32514, USA - zhu@uwf.edu

ISPRS Technical Commission III

KEY WORDS: PM2.5, stroke, remote sensing, health effect, air pollution

ABSTRACT:

This research explores the use of PM2.5 gird derived from remote sensing for assessing the effect of long-term exposure to PM2.5 (ambient air pollution of particulate matter with an aerodynamic diameter of 2.5 μm or less) on stroke, adjusting for unhealthy behaviors and medical risk factors. Health data was obtained from the newly published CDC "500 Cities Project" which provides city- and census tract-level small area estimates for chronic disease risk factors, and clinical preventive service use for the largest 500 cities in the United States. PM2.5 data was acquired from the "The Global Annual PM2.5 Grids from MODIS, MISR and SeaWiFS Aerosol Optical Depth (AOD), V1 (1998-2012)" datasets. Average PM2.5 were calculated for each city using a GIS zonal statistics function. Map data visualization and pattern comparison, univariate linear regression, and a multivariate linear regression model fitted using a generalized linear model via penalized maximum likelihood found that long-term exposure to ambient PM2.5 may increase the risk of stroke. Increasing physical activity, reducing smoking and body weight, enough sleeping, controlling diseases such as blood pressure, coronary heart disease, diabetes, and cholesterol, may mitigate the effect. PM2.5 grids derived from moderate resolution satellite remote sensing imagery may offer a unique opportunity to fill the data gap due to limited ground monitoring at broader scales. The evidence of raised stroke prevalence risk in high PM2.5 areas would support targeting of policy interventions on such areas to reduce pollution levels and protect human health.

1. INTRODUCTION

A stroke is a brain attack which occurs when a blood vessel that carries oxygen and nutrients to the brain is either blocked by a clot (ischemic stroke) or bursts (hemorrhagic stroke) (Cotran, et al., 1970). When a stroke happens, oxygen supply is deprived and brain cells begin to die, causing the patient to lose abilities controlled by the affected area of the brain such as memory and muscle control. Stroke is the fifth leading cause of death in the United States and is a major cause of serious disability for adults (Kochanek, 2014). About 795,000 people in the United States have a stroke each year (Mozzafarian, et al., 2016). Approximately 610,000 of these are first or new strokes. About 185,000 strokes are in people who have had a previous stroke. About 87% of all strokes are ischemic strokes, when blood flow to the brain is blocked (Mozzafarian, et al., 2016). Stroke kills almost 130,000 Americans each year – approximately 1 out of every 20 deaths (CDC, 2015).

Well known stroke risk factors include age, race, sex, use of hormones, lifestyle risk factors (e.g., obesity, physical inactivity, cigarette smoking or exposure to secondhand smoke), and medical risk factors (e.g., high blood pressure, high cholesterol, diabetes, and obstructive sleep apnea). However, the potential effect of ambient air pollution is much less recognized. Recent years have seen more and more studies that investigate the association between short-term or long-term exposure to PM2.5 (ambient air pollution of particulate matter with an aerodynamic diameter of 2.5 μm or less), and stroke mortality (Hu, et al, 2008; Lin, et al., 2017; Matthew, et al., 2013; McClure, et al, 2017). Existing studies often use air pollutant concentration ground monitoring data or spatial proximity to polluters as a surrogate exposure. Some use PM2.5 raster grid derived from remotely sensed imagery. Existing research results are inconsistent and inclusive, especially when multiple risk factors are present. There is a need to continue to investigate the association between stroke and PM2.5. The objective of this study was to examine if there is association between stroke and PM2.5, adjusting for unhealthy behaviors and medical risk factors.

2. DATA

2.1 Health Data

Heath data was obtained from "500 Cities Project" (https://www.cdc.gov/500cities/index.htm). This project reports city and census tract-level data, obtained using small area estimation methods, for 27 chronic disease measures (including unhealthy behaviors, health outcomes, and prevention practices) for the 500 largest American cities. The method of generating small area estimation of the measures is a multi-level statistical modeling framework (Zhang, et al., 2014). The primary data sources for this project are the CDC Behavioral Risk Factor Surveillance System (BRFSS), the Census 2010 population, and the American Community Survey estimates. The project represents a first-of-its kind data analysis to release information on a large scale for cities and for small areas within cities. This system complements existing surveillance data necessary to more fully understand the health issues affecting the residents of that city or census tract. The project includes a total population of 103,020,808, which represents 33.4% of the total United States population of 308,745,538.

* Corresponding author

This study focuses on the conterminous U.S. only. Cities outside conterminous U.S. were excluded from the analysis, leading to a total of 491 cities. Health data from the project for this study includes: (1) health outcomes: stroke, high blood pressure, coronary heart disease, diabetes, high cholesterol, and (2) unhealthy behaviors: smoking, no leisure-time physical activity, obesity, and sleeping less than 7 hours among adults. All data are for adults aged ≥18 years. All data are represented as annual prevalence rate (%) and age-standardized by the direct method to the year 2000 standard U.S. population.

2.2 PM2.5

PM2.5 data for the year 2012 was acquired from the "The Global Annual PM2.5 Grids from MODIS, MISR and SeaWiFS Aerosol Optical Depth (AOD), V1 (1998-2012)" datasets. Annual mean grids of PM2.5 were derived from a combination of MODIS (Moderate Resolution Imaging Spectroradiometer), MISR (Multi-angle Imaging SpectroRadiometer) and SeaWIFS (Sea-Viewing Wide Field-of-View Sensor) AOD satellite retrievals (van Donkelaar, et al., 2015). An annual grid covers the globe providing a continuous PM2.5 (in µg/m3) surface. PM2.5 was derived using the GEOS-Chem chemical transport model which established the local relationships between total column retrievals of AOD and PM2.5. The model appropriately represents temporal trends by considering time variation in the local relationships. The raster grids have a grid cell resolution of 6 arc-minutes (0.1 degree or approximately 10 km at the equator) and cover the global land surface from 70 degrees north to 55 degrees south.

The health data was at two different levels: city-level and census tract-level. The remote sensing data has a resolution of about 10 km while average census tract has an average size of 2.75 by 2.75 km2. Each PM2.5 grid cell on average covers about 16 census tracts. The census tracts within a same pixel have the same PM2.5 values but different health variable values, which makes census tract-level statistical correlation analysis improper. The average city size is 20 by 20 km2 and covers about four pixels of the remote sensing data, therefore city-level analysis makes more statistical sense. Table 1 shows a list of variables.

Variables	Description
	Dependent variable
STROKE	Stroke among adults aged ≥18 years, prevalence rate (%)
	Predictors
PM_{25}	Concentration of particulate matter with an aerodynamic diameter of 2.5 µm or less (µg/m^3)
BPHIGH	High blood pressure among adults aged ≥18 years, prevalence rate (%)
CHD	Coronary heart disease among adults aged ≥18 years, prevalence rate (%)
SMOKING	Current smoking among adults aged ≥18 years, prevalence rate (%)
DIABETES	Diabetes among adults aged ≥18 years, prevalence rate (%)
HIGHCHOL	High cholesterol among adults aged ≥18 years, prevalence rate (%)
LPA	No leisure-time physical activity among adults aged ≥18 years, prevalence rate (%)
OBESITY	Obesity among adults aged ≥18 years, prevalence rate (%)
SLEEP	Sleeping less than 7 hours among adults aged ≥18 years, prevalence rate (%)

Table 1. List of variables

3. DATA VISUALIZATION AND ANALYSIS

3.1 Data Visualization

Figure 1 shows two maps that use graduated color to display and compare the spatial patterns of age-adjusted stroke prevalence rates and the environment factor of interest - PM2.5 concentrations. The stroke map shows higher stroke prevalence in eastern U.S (along the Atlantic coast in the north, in southeastern Georgia, South Carolina, and North Carolina cities, Florida, Gulf coast cities from Louisiana to Texas, as well as in cities near the Great Lakes area) and southern California. The PM2.5 map shows similar pattern except lower concentrations in Gulf coast and Florida, as well as higher values in the eastern part of Midwest.

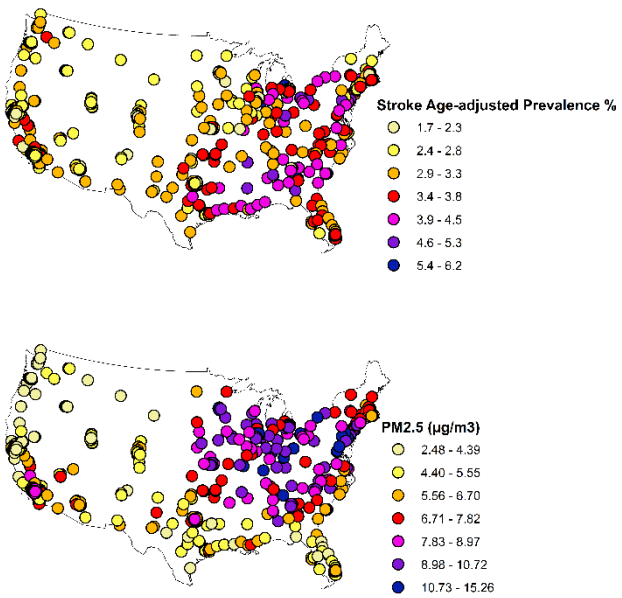

Figure 1. Stroke age-adjusted prevalence (%) vs. PM2.5 (µg/m3), 491 major cities in conterminous USA.

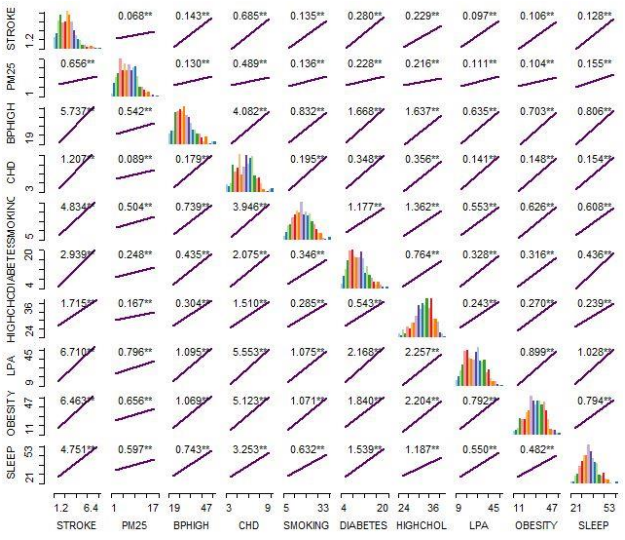

Figure 2. Scatter plots.

Figure 2 shows a scatter plot for each variable-pair of the 10 variables. All the variables approximate to normal distribution. Stroke is positively correlated to all the predictors, but PM2.5 has

relatively lower correlation. It should be noted that PM2.5 and other predictor variables also correlate with each other. This could cause a multi-collinearity issue in a multivariate linear regression model.

3.2 Univariate and Multivariate Linear Regressions

First, a univariate linear regress model was fitted for each predictor. Table 2 shows the results. Stroke prevalence is significantly (at the 99% confidence level) and positively associated with all the predictors, including PM2.5 (p is close to 0). However, R-squared coefficients of determination are not consistent across the variables. PM2.5 has lowest R2 (0.0446). The low R2 value indicates that PM2.5 alone is not sufficient to explain the variance in the stroke prevalence rates. Medical and unhealthy behavioral risk factors must be included in the model. Medical and unhealthy behavioral risk factors have R2 values from 0.3931 (high cholesterol) to 0.8222 (diabetes). The only insignificant intercept is for obesity, although obesity has a significant and positive slope value.

Dependent variable	STROKE				
Predictors	Intercept	*p*	Slope	*p*	R^2
PM_{25}	2.5608	0.0000	0.0680	0.0000	0.0446
BPHIGH	-1.2858	0.0000	0.1426	0.0000	0.8179
CHD	-0.8725	0.0000	0.6354	0.0000	0.8272
SMOKING	0.6543	0.0000	0.1354	0.0000	0.6543
DIABETES	0.1937	0.0018	0.2797	0.0000	0.8222
HIGHCHOL	-4.1366	0.0000	0.2293	0.0000	0.3931
LPA	0.5541	0.0000	0.0967	0.0000	0.6488
OBESITY	-0.0441	0.6498	0.1057	0.0000	0.6830
SLEEP	-1.5245	0.0000	0.1280	0.0000	0.6079

Table 2. Univariate linear regression

Dependent variable	Stroke	
Predictors	Coefficient	*p*
(INTERCEPT)	-0.1347	0.4402
PM_{25}	-0.0087	0.0204
BPHIGH	0.0426	0.0000
CHD	0.2247	0.0000
SMOKING	0.0383	0.0000
DIABETES	0.1611	0.0000
HIGHCHOL	-0.0381	0.0000
LPA	-0.0355	0.0000
OBESITY	0.0077	0.0393
SLEEP	0.0064	0.0702
Model performance and diagnostic		
R^2	0.9451	
Adjusted R_2	0.9441	
F-statistic	920.10	
P (F-statistic)	0.0000	
Multicollinearity condition number	103	

Table 3. Multivariate linear regression

Second, a multivariate linear regression model was fit. The results are shown in Table 3. All predictors are significant at the 90% confidence level. However, slope signs for PM2.5, high blood pressure, coronary heart disease, high cholesterol, obesity, and sleep become negative, contradictory to what the scatter plots show and the results from the univariate regression. The flipping of signs is caused by the collinearity among the predictors. The multi-collinearity condition number is 103, far exceeding the acceptable criterion value of 30, indicating extremely high collinearity. Predictors with relatively lower correlation were suppressed by the ones with higher correlations in the competition for explaining the variation in stroke rates. Although the model has a very high R2 value (>94%), the collinearity has skewed the relationship between some of the predictors and stroke. The associations derived from this model are not reliable. A multivariate model that can deal with collinearity is needed if all potential risk factors need to be kept in the model.

3.3 Linear Regression via Penalized Maximum Likelihood

To overcome the multi-collinearity issue, a linear regression was fitted using a generalized linear model (GLM) via penalized maximum likelihood using the glmnet R package (Friedman et al., 2010). The package computes the regularization path for the lasso or elasticnet penalty at a grid of values for the regularization parameter (λ). Gaussian family type was assumed for the dependent variable STROKE.

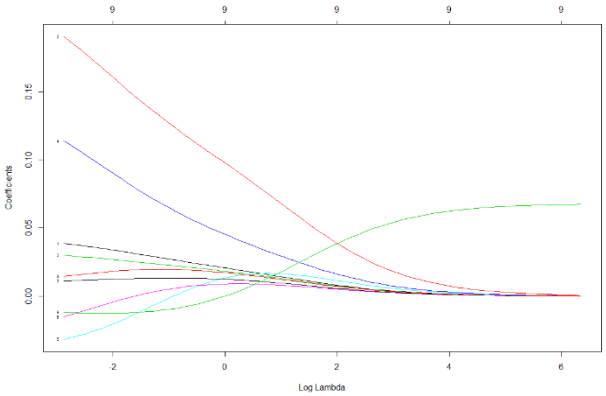

Figure 3. Regularization path for the elasticnet penalty. Each curve corresponds to a predictor variable. It shows the path of its coefficient against log lambda (lambda is a complexity parameter: λ≥0). 1 – BPHIGH, 2-CHD, 3-SMOKING, 4-DIABETES, 5-HIGHCHOL, 6-LPA, 7-OBESITY, 8-SLEEP, 9-PM25.

Dependent variable	Stroke
Predictors	Coefficient
(INTERCEPT)	0.9140
PM_{25}	0.0346
BPHIGH	0.0091
CHD	0.0442
SMOKING	0.0085
DIABETES	0.0185
HIGHCHOL	0.0127
LPA	0.0056
OBESITY	0.0065
SLEEP	0.0080

Table 4. Ridge regression results (logλ= 1.8, λ = 6.05, α=0)

The elasticnet mixing parameter alpha was set to 0 (ridge penalty). The ridge penalty shrinks the coefficients of correlated predictors towards each other. The number of lambda values was set to 100. The predictors were standardized. A penalty factor of 0 was applied to the variable of interest (PM2.5). A value of 0 for the penalty factor implies no shrinkage. The sequence of models implied by lambda was fit by coordinate descent. For the Gaussian family, this is the elasticnet sequence for alpha = 0.

Figure 3 shows the regularization path for the elasticnet penalty at a grid of values for the regularization parameter (λ) for each of the predictor variables. A cross-validation function in the glmnet

package was run to select a lambda value. The function is a predictive criterion that evaluates the sample performance by splitting the sample into training and validation sets and choosing the value of lambda with which the error of prediction is minimal. A lambda value of 6.05 (log λ= 1.8) was selected which meets the criterion. Coefficients values from the model at the lambda value are shown in Table 4. All predictors show positive association with stroke.

4. CONCLUSION

Ambient PM2.5 may increase the risk of stroke. Increasing physical activity, reducing smoking and body weight, adequate sleep, and controlling diseases such as blood pressure, coronary heart disease, diabetes, and cholesterol, may mitigate the effect. With its very fine particle size, PM2.5 can penetrate deeply the human lung, enter the blood system, and circulate into the central nervous system. PM2.5 can trigger inflammatory processes and cause chronic pulmonary and systemic oxidative stress, leading to a series of vasculature and blood tissue level processes critical for the development of stroke (Scheers, et al., 2015). PM2.5 grids created by moderate resolution satellite remote sensing imagery may offer a unique opportunity to fill the data gap due to limited ground monitoring at broader scales. The evidence of raised stroke prevalence risk in high PM2.5 areas would support targeting of policy interventions on such areas to reduce pollution levels and protect human health.

REFERENCES

CDC, 2015. Underlying cause of death 1999-2013 on CDC WONDER Online Database, released 2015. Data are from the Multiple Cause of Death Files, 1999-2013, as compiled from data provided by the 57 vital statistics jurisdictions through the Vital Statistics Cooperative Program. Accessed January 8, 2018.

Cotran, R.S., Kumar, V., Fausto, N., Robbins, S.L., Abbas, A.K., 1970. Robbins and Cotran pathologic basis of disease. St. Louis, Mo: Elsevier Saunders.

Friedman, J.H., Hastie,T., Tibshirani, R., 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1), 1-22.

Hu, Z., Liebens, J., Rao, R.K., 2008. Linking stroke mortality with air pollution, income, and greenness in northwest Florida: an ecological geographical study. *International Journal of Health Geographics* 7(20), https://doi.org/10.1186/1476-072X-7-20.

Kochanek, K.D., Xu, J.Q., Murphy,S.L., Arias, E.,2014. Mortality in the United States, 2013. NCHS Data Brief, No. 178. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention, Department of Health and Human Services.

Lin, H., Guo, Y., Di, Q., Zheng, Y., Kowal, P., Xiao, J., Liu, T., Li, X., Zeng, W., Howard, S.W., Nelson, E.J., Qian, Z., Ma, W., Wu, F., 2017. PM2.5 and stroke: effect modifiers and population attributable risk in six low- and middle-income countries. *Stroke* 48:1191-1197. https://doi.org/10.1161/STROKEAHA.116.015739.

Matthew, S.L., Kent, S.T., Al-Hamdan, M.Z., Crosson,W.L., Estes, S.M., Estes Jr, M.G., Quattrochi, D.A., Hemmings, S.N., Wadley, V.G., McClure, L.A., 2013. Fine particulate matter and incident cognitive impairment in the REasons for geographic and racial differences in stroke (REGARDS) cohort. *PLOS One* 8(9): e75001. https://doi.org/10.1371/journal.pone.007500.

McClure, L.A., MS, L.P., Crosson, W., Kleindorfer, D., Kissela, B., Al-Hamdan, M. 2017. Fine Particulate Matter (PM2.5) and the Risk of Stroke in the REGARDS Cohort. *Journal of Stroke and Cerebrovascular Diseases* 26(8):1739-1744. doi:10.1016/j.jstrokecerebrovasdis.2017.03.041.

Mozzafarian, D., Benjamin, E.J., Go, A.S., Arnett, D.K., Blaha, M.J., Cushman, M. 2016. Heart disease and stroke statistics—2016 update: a report from the American Heart Association. *Circulation* 133(4):e38–360.

Scheers, H., Jacobs, L., Casas, L., Nemery, B., Nawrot, T.S., 2015. Long-term exposure to particulate matter air pollution is a risk factor for stroke: meta-analytical evidence. *Stroke* 46:3058–3066. doi: 10.1161/STROKEAHA.115.009913.

Zhang, X., Holt, J.B., Lu, H., Wheaton, A.G., Ford, E.S., Greenlund, K.J., Croft, J. B., 2014. Multilevel regression and poststratification for small-area estimation of population health outcomes: a case study of chronic obstructive pulmonary disease prevalence using the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology* 179(8), 1025-1033, DOI: 10.1093/aje/kwu018.

van Donkelaar, A., Martin, R.V., Brauer, M., Boys, B. L., 2015. Global Annual PM2.5 Grids from MODIS, MISR and SeaWiFS Aerosol Optical Depth (AOD), 1998-2012. Palisades, NY: NASA Socioeconomic Data and Applications Center (SEDAC). http://dx.doi.org/10.7927/H4028PFS. Accessed March 15, 2018.