Hygiene and Environmental Health Advances 4 (2022) 100039

Contents lists available at ScienceDirect

Hygiene and Environmental Health Advances

journal homepage: www.elsevier.com/locate/heha

How neighborhood environment modified the effects of power outages on multiple health outcomes in New York state?

Wangjian Zhang[a,b], Xinlei Deng[b], Xiaobo X. Romeiko[b], Kai Zhang[b], Scott C. Sheridan[c], Jerald Brotzge[d], Howard H. Chang[e], Eric K. Stern[f], Zhijian Guo[g], Guanghui Dong[h], Ramune Reliene[i], Yuantao Hao[a,*], Shao Lin[b,*]

[a] Department of Medical Statistics, School of Public Health, Sun Yat-sen University, Guangzhou, China
[b] Department of Environmental Health Sciences, University at Albany, State University of New York, Rensselaer, NY, USA
[c] Department of Geography, Kent State University, Kent, OH, USA
[d] New York State Mesonet, College of Arts and Sciences, State University of New York, Rensselaer, NY, USA
[e] Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA
[f] College of Emergency Preparedness, Homeland Security and Cybersecurity, State University of New York, Albany, NY, USA
[g] Department of Mathematics and Statistics, State University of New York, Albany, NY, USA
[h] Department of Preventive Medicine, School of Public Health, Sun Yat-sen University, Guangzhou, China
[i] Cancer Research Center, State University of New York, Rensselaer, NY, USA

ARTICLE INFO

Key words:
Power outage
Neighborhood environment
Multiple health outcomes
Vulnerability
Machine learning

ABSTRACT

Background: Although power outage (PO) is one of the most important consequences of increasing weather extremes and the health impact of POs has been reported previously, studies on the neighborhood environment underlying the population vulnerability in such situations are limited. This study aimed to identify dominant neighborhood environmental predictors which modified the impact of POs on multiple health outcomes in New York State.

Methods: We applied a two-stage approach. In the first stage, we used time series analysis to determine the impact of POs (versus non-PO periods) on multiple health outcomes in each power operating division in New York State, 2001-2013. In the second stage, we classified divisions as risk-elevated and non-elevated, then developed predictive models for the elevation status based on 36 neighborhood environmental factors using random forest and gradient boosted trees.

Results: Consistent across different outcomes, we found predictors representing greater urbanization, particularly, the proportion of residents having access to public transportation (importance ranging from 4.9–15.6%), population density (3.3–16.1%), per capita income (2.3–10.7%), and the density of public infrastructure (0.8–8.5%), were associated with a higher possibility of risk elevation following power outages. Additionally, the percent of minority (-6.3–27.9%) and those with limited English (2.2–8.1%), the percent of sandy soil (6.5–11.8%), and average soil temperature (3.0–15.7%) were also dominant predictors for multiple outcomes. Spatial hotspots of vulnerability generally were located surrounding New York City and in the northwest, the pattern of which was consistent with socioeconomic status.

Conclusion: Population vulnerability during power outages was dominated by neighborhood environmental factors representing greater urbanization.

Introduction

Power outages (POs) are one of the most common consequences of extreme weather events worldwide, and one that may worsen in the context of increased energy demand as well as climate change. For example, Hurricane Maria in 2017 left households in Puerto Rico out of power for 84 days on average (Kishore et al., 2018a), resulting a total of 4.0 billion customer hours of interruption (Román et al., 2019). The heavy dependence of modern infrastructure on electricity can lead to significant public health impacts when power is lost (Anderson and Bell, 2012a; Christine et al., 2019a; Kishore et al., 2018b).

Abbreviations: PO, Power outage.

* Corresponding authors.
E-mail addresses: haoyt@mail.sysu.edu.cn (Y. Hao), slin@albany.edu (S. Lin).

https://doi.org/10.1016/j.heha.2022.100039
Received 5 May 2022; Received in revised form 16 November 2022; Accepted 18 November 2022

A few studies have provided epidemiological evidence on the adverse health impact of power outages. For example, Christine et al (Christine et al., 2019b) reported significant elevations in all-cause mortality and hospitalizations for multiple health outcomes due to citywide and localized power outages in New York City. Lin et al(Lin et al., 2011a) observed that mortality and respiratory hospital admissions following the August 14, 2003, Northeast Blackout were significantly elevated by two to eight fold. Most recently, we observed significant associations of power outages with the elevated rate of COPD hospitalization as well as increased severity of symptoms and hospital charges (Zhang et al., 2020b).

These evidences provided critical quantitative supporting information for future preventive strategies designed to improve the electricity supply system, however, significant gaps remain in our further understanding of how these adverse health impacts of power outages were shaped by the neighborhood environmental characteristics. Some studies suggested that lower socioeconomic characteristics such as a higher low-education rate, a higher poverty rate and a higher percentage of minorities were associated with a higher health impact of general environmental hazards(Barry E et al., 2011a; Nayak et al., 2020; Zoraster, 2010). One of our previous studies suggested that housing and transportation factors such as a higher percentage of people living in multi-unit structures or group quarters were top contributors to an increased impact of catastrophic storms on cardiovascular diseases and mental disorders (Zhang et al., 2020a). Identification of such dominant predictors is critical for the development of preventive strategies targeted on communities with certain characteristics or at locations. However, no studies have ever been designed to examine the neighborhood environmental predictors underlying the health impact of power outages.

To address this important knowledge gap, we conducted a large population-based study in New York State to identify dominant neighborhood environmental predictors including socioeconomic indicators, landscape characteristics and built environment metrics. We also developed predictive models for the health impact of power outages based on these predictors and assessed the population vulnerability in order to identify the geographical hotspots. Given the numerous predictors involved in the single model, we used sophisticated machine-learning methods to address the potential collinearity issue as well as to accommodate the complex relationships between the outcome and predictors and inter-predictor interactions.

Materials and methods

Study population and design

This study covered the entire population of New York State. All hospital admission records with a principal diagnosis of cardiovascular diseases, respiratory diseases, respiratory infectious diseases, food-and-water-borne diseases or injuries between 2001 and 2013 were included. We focused on these outcomes since they were the major disease groupings previously reported to be associated with environmental stressors and were biologically plausible (Anderson et al., 2013; Anderson and Bell, 2012b; Bloom et al., 2016; Christine et al., 2019b; Li et al., 2019).

As described in Fig. 1, we first conducted a time-series analysis to examine whether the risk ratio (RR) of a health outcome (associated with power outages) increased (RR>1) following POs at each power operating division. A power operating division is a spatial unit in which an electricity company operates and maintains the electric distribution facilities. The entire state contains ~1700 power operating divisions, with an average population of 11,061 per division. We then linked the risk elevation status with relevant predictors at the division level and developed predictive models based on machine learning algorithms. We finally used the optimal predictive model to identify the dominant predictors and the spatial pattern of the health impact of power outages. With the time-series analysis, we were able to control important time-varying confounders and capture the cumulative health effect of power outages (Bhaskaran et al., 2013a; Gasparrinia et al., 2010). The machine learning methods have multiple unique advantages including accommodating correlations between predictors and outlier issues in the data and having better predictive performance than traditional parametric models (Zhang et al., 2016a).

Health data and outcomes

We retrieved the hospital admission data between 1/1/2001-12/31/2013 from the New York Statewide Planning and Research Cooperative System (SPARCS, https://www.health.ny.gov/ statistics/sparcs/), a legislatively mandated database covering 95% of hospitals across the state (Rich et al., 2019; Zhang et al., 2018). We retained records with a principal diagnosis of major population health concerns including cardiovascular diseases (the International Classification of Diseases, Ninth Revision (ICD 9) code: 393-396, 401-405, 410-415, 427, 428, 430-434, 436-438), respiratory diseases (ICD 9 code: 480-488, 491-496, 518), respiratory infectious diseases (ICD 9 code: 480-488), food-and-water-borne diseases (ICD 9 code: 001-009) or injuries (ICD 9 code: E880-E910). We geocoded each hospitalization record to the power operating division level based on the residential address reported to the SPARCS. We defined the outcome for the time-series analysis as the daily number of hospital admissions in each power operating division and for each health outcome. The outcome for the machine learning methods was the risk elevation status (0/1) identified based on the results of the time-series analysis.

Exposure data and predictors

For each operating division, we obtained the total number of customers, the date a power outage occurred as well as the number of customers affected from the NYS Department of Public Service. We divided the number of affected customers over the total number of customers to calculate the coverage of power outage for each day in each division. We identified the 50^{th} percentile of the coverage among all PO days in all divisions which was 0.5% and defined a day with PO coverage above this cutoff as an exposure day, otherwise as a control day. This criterion was selected based on previous studies and to capture the health impact of small localized power outage events(Sheridan et al., 2021; Zhang et al., 2020c).

To develop the predictive model for the health impact of power outage, we retrieved data on predictors including socioeconomic variables from the American Community Survey (ACS) (Barry E et al., 2011b), landscape variables from the National Land Coverage Database (NLCD) (Homer et al., 2015) and Soil Survey Geographic Database (SSURGO) (Bocinsky, 2019), and built environment variables from the Environmental Quality Index (EQI) of the U.S. Environmental Protection Agency (EPA) (Lobdell et al., 2011). A full list of these neighborhood environmental factors are presented in the results section. We included these factors as the major predictors since they were considered to be the most important and comprehensive predictors for residential environment and population vulnerability (Barry E et al., 2011c; GISP, n.d.; Nayak et al., 2018; Zhang et al., 2020a).

Statistical analysis

Divisions with < 100 residents or < 5 cases per year were excluded to ensure a reasonable sample size, leaving 407-1574 divisions included (accounting for 80.0–93.3% of the total cases for different outcomes). *In the first stage*, we developed a distributed lag nonlinear models (DLNM) with quasi-Poisson distribution in each division for each health outcome (Bhaskaran et al., 2013b; Gasparrini, 2013). Specifically, we regressed the daily number of cases against the indicator of exposure/control days, meanwhile adjusting for time-varying confounders including day

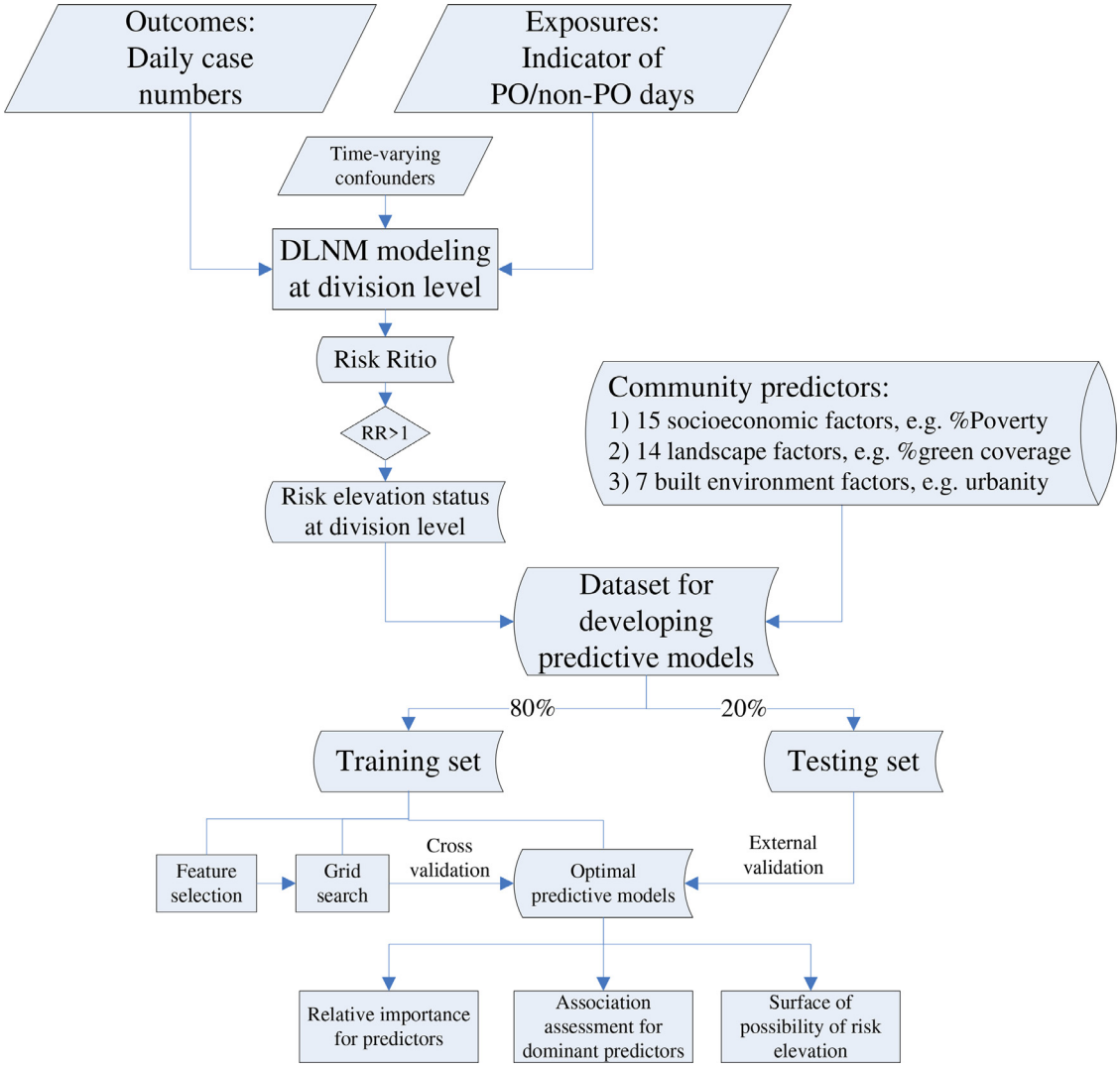

Fig. 1. Diagram of Model Development and Application. The health risk elevation status identified using the time-series analysis was linked to a comprehensive set of 36 factors at the community level to develop the predictive models and to evaluate the contribution of each factor to the health risk elevations.

of the week, holidays, the long-term trend and seasonality, and weather confounders including temperature and humidity (Xiao et al., 2017; Zhang et al., 2016b). We also controlled the ambient concentration of $PM_{2.5}$, the most important air pollutant, and weather events including hurricanes, winter storms, thunderstorms, flooding, strong wind events and heat events as reported to NOAA (ftp://ftp.ncdc.noaa.gov). In this stage, we estimated the cumulative health effect (i.e., risk ratio, *RR*) within the first week (0-6 lag days) following POs and defined divisions with *RR* >1 as risk elevated communities (i.e., 1), all others as non-elevated communities (i.e., 0).

In the second stage, with the dataset containing the risk elevation status (0/1) variable and numerous predictors at the community (i.e., division) level, we developed and applied predictive models in following steps (LeDell et al., 2019a):

1) *Dataset splitting.* We generated a random variable, included it into the dataset and split the dataset into the training set (80% records) and the testing set (20% records).
2) *Feature selection.* As there is not gold standard for an optimal model used for the initial step of variable selection, we ran a gradient boosted model (additionally with an initial learning rate of 0.01) or a random forest model using 200 trees, 3 folds of cross validation, a maximum interactive depth of 5, and a sampling rate of 0.7. We utilized two major tree models with which we focused on the relative importance of predictors, and selected those with an importance greater than the random term.
3) *Grid search.* We refitted predictive models based on selected predictors. To determine the model with the best predictive performance, we searched for the optimal combination of parameters (i.e., hyper-parameter) based on cross validation with the grid search feature of the h2o package in R. More details regarding the parameters in grid search are presented in Table 1. We also added a built-in balance procedure or SMOTE (Synthetic Minority Over-sampling Technique) procedure to the original models to balance the class distribution and to check the improvement in the model performance (Torgo, 2010a).
4) *External validation and application.* We validated the optimal predictive model with the testing set. With the optimal model, we determined the relative importance of each predictor, identified the dominant predictors, and investigated the association of those predictors with the health impacts of power outages. Finally, we utilized the full predictor data at the community (i.e., division) level to predict the possibility of elevated health risk following POs across the entire state covering the less populated divisions (accounting for 16.7–20% cases) where statistical modeling cannot be developed.

Table 1
Grid Search and Optimal Model Specifications.

Hyperparameters	Balance strategies	Optimal models
Gradient boosted models (GBM): Number of trees (ntrees): 10-200 by 10; 200-500 by 50 Maximum depth of interaction (max_depth): 3-5 Learning rate: 0.001, 0.005, 0.01, 0.1	None Built-in balance procedure	CVD: RF with SMOTE, with ntrees=500 and max_depth=5 Respiratory: RF with SMOTE, with ntrees=350 and max_depth=3 Respiratory infections: RF with SMOTE, with ntrees=60 and max_depth=4
Random forest (RF): Number of trees: 10-200 by 10; 200-500 by 50 Maximum depth of interaction: 3-5	SMOTE (Synthetic Minority Over-sampling Technique)	FWBD: RF with SMOTE, with ntrees=20 and max_depth=3 Injuries: RF with SMOTE, with ntrees=40 and max_depth=4

We completed geocoding using the Street and Address Maintenance Program in ArcGIS 10.3.1 (The NYS GIS Program Office, 2017) and accomplished all analyses with R 3.4.1.

Results

Descriptive statistics

There was a total of 3,537,802 hospital admissions due to cardiovascular diseases, 1,906,429 due to respiratory diseases, 808,755 due to respiratory infections, 166,772 due to food and water borne diseases and 1,037,245 due to injuries reported across the New York State during the study period. We observed that power outages overall were associated with an increased hospitalization rate in 29.7% electric operating divisions for cardiovascular diseases, 24.2% for respiratory diseases, 20.1% for respiratory infectious diseases, 21.9% for food and water borne diseases, and 29.1% for injuries, as compared with non-power outage periods.

Overview of the predictor contributions

As described in Table 1, the optimal models we identified for those outcomes generally were models based on the random forest algorithm with outcomes balanced using the SMOTE procedure. The optimal number of trees and the optimal maximum depth of interactions varied across health outcomes. Based on the optimal predictive models, we found the average relative importance of predictors were generally consistent across the three categories: the socioeconomic factors (ranging 3.5–5.4% for outcomes except food and water borne diseases), landscape factors (ranging 1.6–4.6%), and built environment factors (ranging 4.3–5.3% for outcomes except food and water borne diseases). The relative importance of socioeconomic factors was slightly higher for respiratory infections whereas built environment factors were slightly more important for the cardiovascular and respiratory diseases compared with other outcomes, as described in Table 2. The average importance of predictors for the food and water borne diseases generally was larger as fewer predictors were identified.

Dominant predictors in different categories

When comparing different factors within each category, we observed that 1) within the socioeconomic category factors representing a greater urbanization such as population density (importance ranging 3.3–16.1%) and per capita income (importance ranging 2.3–10.7%) generally were associated with a higher possibility of elevated health risk during power outages. In contrast, less urbanized divisions such as those with a higher percent of residents living in mobile homes (importance ranging -10.2% to -2.9%) usually had a lower health risk. The percent of minority, population density, per capita income, and percent of mobile homes were top contributors to the vulnerability of multiple health outcomes as displayed in Fig. 3. 2) In addition, we found that elevated health risk was also associated with other urbanization indicators such as the proportion of residents having access to public transportation (importance ranging 4.9–15.6%) and the density of public infrastructure including healthcare (importance ranging 3.9–8.5%) and education (importance ranging 0.8–7.6%) related business. 3) The average soil temperature (importance ranging 3.0–15.7%) and the percent of sand in soil (importance ranging 1.6–11.8%) were two most important landscape predictors for multiple health outcomes, as described in Table 2 and Fig. 3.

Dominant predictors between different outcomes

When comparing different health outcomes, although the overall trend of a higher health risk in divisions with a higher degree of urbanization was generally consistent, top contributors were slightly different for different outcomes. In addition to the common dominant predictors, each health outcome also had its unique important contributors as described in Fig. 3. The risk of cardiovascular diseases and respiratory infections was higher in divisions with a higher percent of sand in soil (importance 11.8% and 6.5%, respectively). The elevated risk of food and water borne diseases was associated with a higher percent of residents with limited English proficiency (importance 8.1%). All health risks were significantly associated with the percent of minority among the total population (importance ranging -6.3% to 27.9%).

Optimal predictive model-based vulnerability of elevated health risk and the spatial pattern

According to Fig. 2, the best predictive model for the hospitalization rate of cardiovascular diseases yielded an area under the receiver operating characteristic curve (AUC) of 0.81 in the model training, and an AUC of 0.80 in the internal cross validation and an AUC of 0.77 in the final external testing. AUCs were generally consistent across training, cross validation and testing for a predictive model which ranged 0.72-0.73 for respiratory diseases, 0.78-0.80 for respiratory infections, 0.77-0.86 for food and water borne diseases, and 0.76-0.83 for injuries. Fig. 4 displayed the predicted possibility surface of risk elevation for each health outcome across the state. Consistent across different health outcomes, elevated risk generally was more likely to occur among the communities in downstate areas surrounding the New York City and among those in the northwest counties including Erie, Monroe and Onondaga.

Discussion

Identifying dominant neighborhood environmental indicators

Higher risk in more urbanized areas. We found that the health impact of power outages overall tended to be higher in areas with a higher degree of urbanization as indicated by a higher population density, greater accessibility to public transportation system, and a higher density of public infrastructure, which was consistent across different health outcomes. This suggests that highly developed and generally reliable urban critical infrastructure systems may be associated with heightened vulnerability,

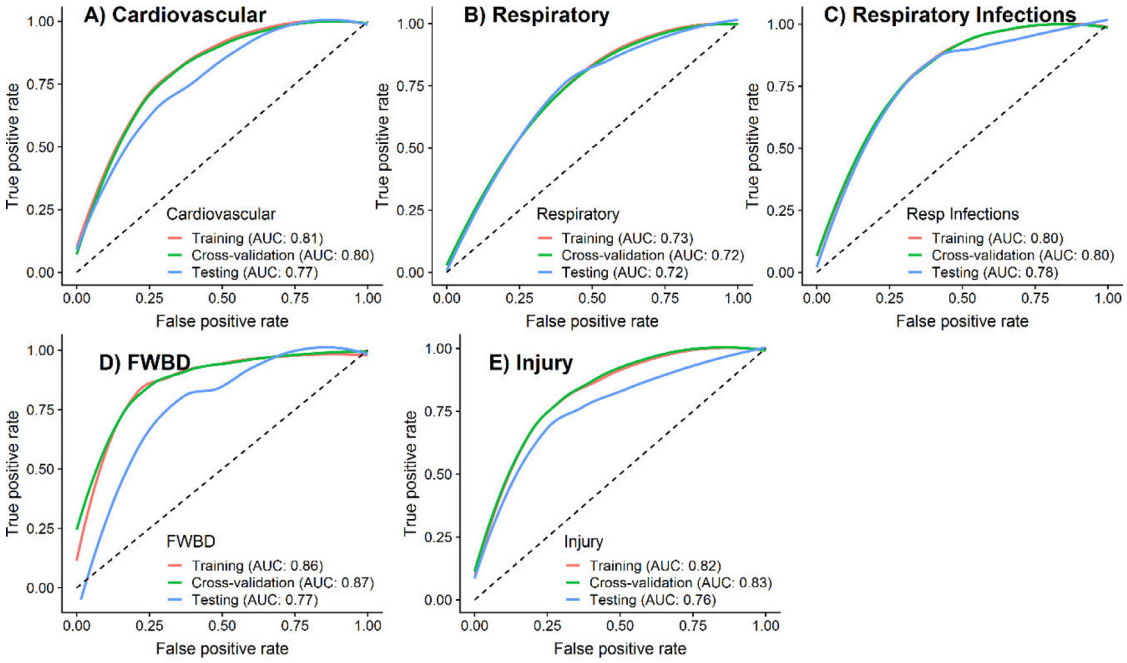

Fig. 2. The Predictive Performance of the Optimal Models. The tuned models were confirmed with internal training, cross-validation and external testing. The area under the ROC curve was generally high, indicating a good predictive performance.

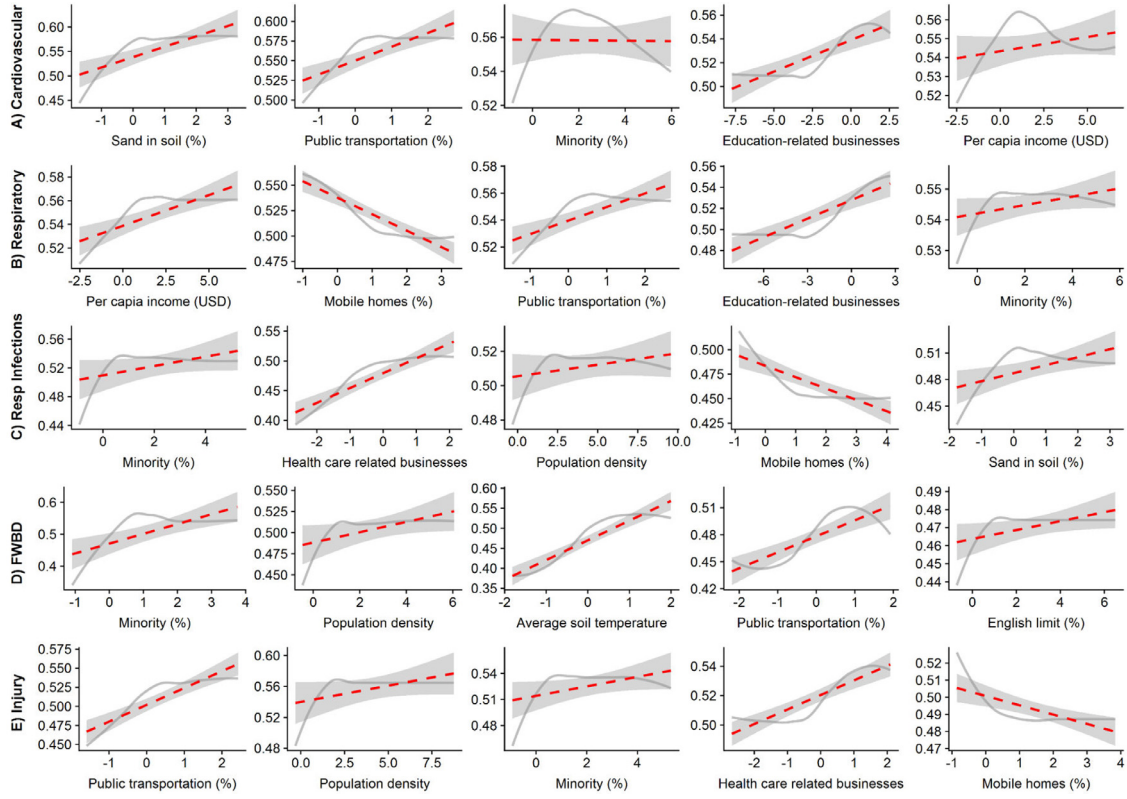

Fig. 3. Associations between Top Five Contributors and the Health Impact of Power Outages, by Health Outcomes. The solid curves represented the marginal effect of a contributor on the outcome while the dashed lines represented the average marginal effect over the range of the contributor as was fitted linearly. Factors representing a greater urbanization such as population density and per capita income generally were associated with a higher possibility of elevated health risk during power outages. In contrast, less urbanized divisions such as those with a higher percent of residents living in mobile homes usually had a lower health risk. The percent of minority, population density, per capita income, and percent of mobile homes were top contributors to the vulnerability of multiple health outcomes.

Table 2
The Relative Importance of Predictors in Predicting the Risk of Hospitalization for Multiple Health Outcomes associated with Power Outages.

Predictors at Community Level	Cardiovascular	Respiratory	Respiratory Infections	FWBD	Injury
Socioeconomic factors					
Under poverty (%)	-3.4	1.8	(D)*	(D)	-3.2
Unemployed (%)	2.4	-1.2	-3.1	(D)	-0.8
Per capita income (USD)	5.2	10.7	2.3	(D)	2.8
Low educated (%)	(D)	1.0	(D)	(D)	-1.3
Elderly (%)	(D)	1.5	3.0	-5.4	0.6
Single-parent households (%)	(D)	(D)	(D)	(D)	-3.4
Minority (%)	-6.3	5.6	21.9	27.9	11.5
Limited English proficiency (%)	2.7	(D)	2.8	8.1	2.2
Multi-unit structure (%)	(D)	(D)	(D)	(D)	(D)
Mobile homes (%)	-2.9	-10.2	-6.5	(D)	-6.0
Living crowded (%)	2.7	(D)	-1.3	(D)	-2.2
No vehicle (%)	-2.9	(D)	(D)	(D)	-1.8
Living in group quarters (%)	(D)	-1.3	(D)	(D)	(D)
Population density (/mile^2)	4.5	3.3	6.8	16.1	14.9
High intensity developed area (%)	2.4	(D)	1.1	-3.2	1.2
Average absolute relative importance	**3.5**	**3.7**	**5.4**	**12.1**	**4.0**
Landscape factors					
Forest (%)	(D)	2.1	(D)	(D)	2.2
Cultivated land (%)	(D)	-2.4	3.2	(D)	-1.6
Wetlands (%)	2.7	-2.0	1.2	(D)	1.8
Open Water (%)	-2.6	3.1	1.5	(D)	-1.2
Grazing land (%)	-2.8	-4.9	-3.6	(D)	-0.5
Sand in soil (%)	11.8	3.3	6.5	2.5	1.6
Silt in soil (%)	-3.1	-1.1	-3.0	(D)	1.3
Clay in soil (%)	-2.0	-1.7	(D)	(D)	-1.4
Organic content in soil (%)	4.5	5.1	1.7	2.7	0.8
Elevation (m)	-2.6	-2.5	-5.4	-1.0	-3.0
Annual average soil temperature (F)	4.6	3.0	3.2	15.7	3.6
Soil temperature variation (F)	2.6	(D)	(D)	3.8	1.2
Annual average soil moisture (%)	-2.0	(D)	(D)	-1.8	-1.0
Soil moisture variation (%)	(D)	0.7	(D)	(D)	-0.9
Average absolute relative importance	**3.8**	**2.7**	**3.3**	**4.6**	**1.6**
Built environment factors					
Urbanicity	(D)	(D)	(D)	(D)	(D)
Public transportation (%)	7.9	10.2	4.9	11.6	15.6
Highway (%)	(D)	4.0	2.3	(D)	-1.1
Primary street (%)	3.6	-2.0	2.3	(D)	0.9
Education-related businesses (/person)	5.6	7.6	3.8	(D)	0.8
Entertainment-related businesses (/person)	2.3	2.3	(D)	(D)	-0.8
Healthcare related businesses (/person)	3.9	5.4	8.5	(D)	6.4
Average absolute relative importance	**4.7**	**5.3**	**4.4**	**11.6**	**4.3**

* (D), excluded in feature selection due to smaller contributions to the outcome compared with the random term.

compared to more rural areas, in the event of disruption (Cutter et al., 2016). Although few studies have evaluated the sociodemographic variations in the health impact of power outages, our finding was in agreement with the limited existing epidemiological evidence. For example, it was reported that respiratory admissions increased by 23% (95%CI: 3–46%) during the 2003 Northeast Blackout in high-income and more urbanized areas of New York City whereas it did not increase in less urbanized regions, as compared with non-blackout days (Lin et al., 2011b). Similar disparities have been reported for the health impact of natural disasters. A long-term study on the hurricane damage on the U.S. Gulf Coast surprisingly revealed that residents at the highest health risk were Whites and nonpoor households (Logan and Xu, 2015). A greater vulnerability among residents in more urbanized areas was expected, particularly during power outages. Electricity is one of the most important components in the process of urbanization. A higher degree of urbanization implies a stronger reliance on electricity (Kaur and Luthra, 2018; Zhao and Zhang, 2018). For example, urban areas are larger consumers of electricity in order to support more public infrastructures and business such as healthcare facilities and water supplies, more home appliances such as air conditioners and humidifiers, and more healthcare devices such as nebulizers and oxygen monitors, compared with less developed areas. Therefore, residents in more urbanized areas generally adapt to a greater electrical dependency and tend to be affected more by power outages compared with those living elsewhere (Zhang et al., 2020a).

Higher risk in areas with higher minorities. Our findings also suggested a higher health impact from power outages in areas with a higher proportion of minorities and those with language barriers. The impact of the percentages of minority and those with language barriers was consistent with numerous previous findings (Di et al., 2017; Parker et al., 2018). These residents are considered the most vulnerable groups in response to environmental stressors, as potential results of obstacles to receiving safety information and help-seeking (e.g. undocumented persons) and/or inadequate access to primary and specific health care and limited health insurance and medical benefits (Bolin and Kurtz, 2018; Muncan, 2018). Our findings also emphasized the importance of potential research and intervention strategies targeting minority communities to improve the health disparities in numerous health issues.

Higher risk with certain soil conditions. Surprisingly, we found a higher health risk of power outages in areas with higher percentages of sandy soil and in areas with a higher soil temperature. In general, sandy soils are lighter and looser than other soil types and have limited water-holding capacity (Fang and Su, 2019; Hagyó and Tóth, 2018). The holding capacity of sandy soils is significantly lower than clay, silt or heavier soils (Chiras, 2009). Therefore, it was likely that the electricity poles and distribution lines were less sustainable in areas with a

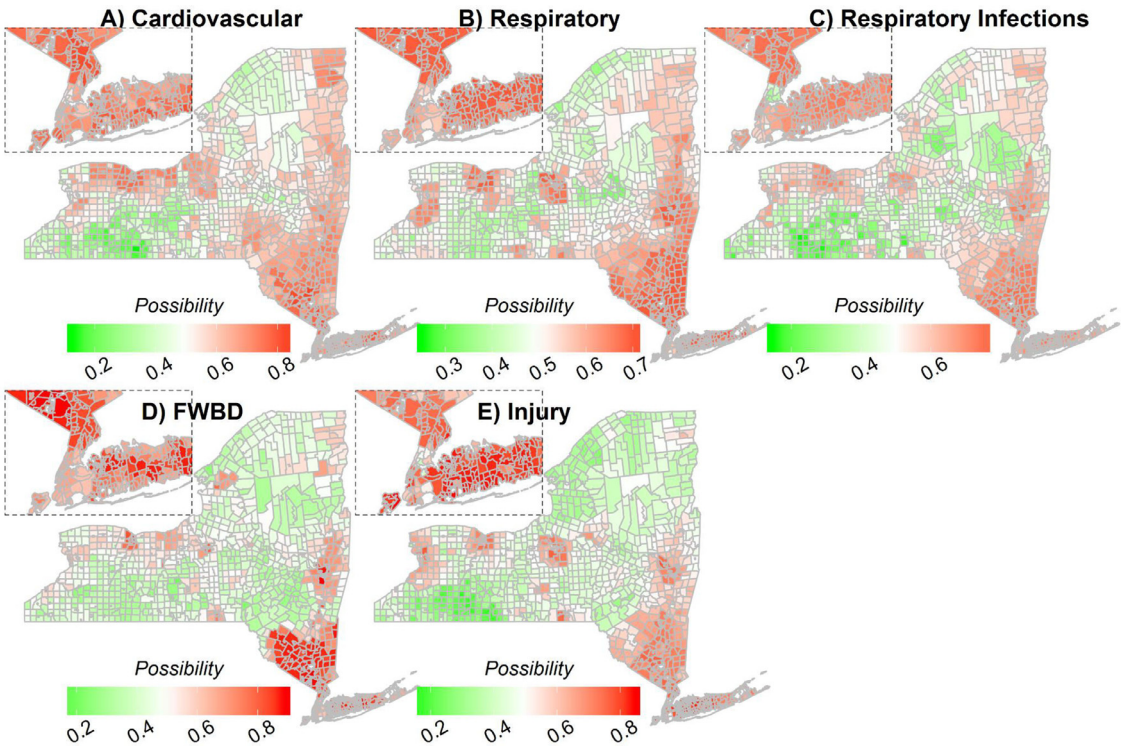

Fig. 4. Predicted Possibility of Elevated Health Risk Associated with Power Outages across the New York State. Consistent across different health outcomes, elevated risk generally was higher in downstate areas and northwest counties including Erie, Monroe and Onondaga.

higher percent of sandy soils. Potential mechanism for the impact of soil temperature remains unknown. However, a possible interpretation was that soil temperature could affect certain biological processes underground, thus, affect the water-retain capacity and holding power of the soil (Nishar et al., 2017). Another possibility was that higher soil temperature may also be associated with higher-intensity human activities (e.g. heavier traffics and heat island effect) in more urbanized regions.

Identifying spatial hotspots with optimal models

According to the optimal predictive models, we identified that the communities in and surrounding the New York City and those located in the northwest counties such as Erie, Monroe and Onondaga were more likely to be the spatial hotspots during power outages. This pattern was consistent across different health outcomes. The potential mechanism underlying the spatial pattern remains unclear. However, it was likely shaped by the joint effect of dominant predictors which were associated with a higher level of urbanization. Specifically, we found that the spatial pattern of the health risk identified in the current study is consistent with the spatial distribution of average wages across the state as reported by the U.S. Bureau of Labor Statistics (New York–New Jersey Information Office, 2016) as well as with the distribution of per capita income reflected by the 2006-2010 American Community Survey 5-Year Estimates (Wikipedia contributors, 2019). These findings confirm that health risks following power outages are generally greater in more urbanized areas than elsewhere, and our models have good predictive capacity to capture these trends.

Optimal predictive models

We found that our models based on machine learning algorithms had good predictive performance in identifying the elevated health risk following power outages. While traditional regression models may suffer from collinearity issues when multiple predictors with high correlation are included in the same model and misfit issues when the impact of a predictor on the outcome is nonlinear, as well as prone to unreliable estimates in presence of outliers in the data, the machine learning methods used in the current study overcome these limitations through a recursive binary split algorithm (Sidey-Gibbons and Sidey-Gibbons, 2019; Tian et al., 2019). The marginal effect of a predictor, regardless of being linear or nonlinear, correlated to the impact of other predictors or not, could be well captured with the machine learning algorithm (Elith et al., 2008; LeDell et al., 2019b). Therefore, machine learning methods usually have a better predictive performance than the traditional methods and have been increasingly used in environmental health studies. Particularly, multiple strategies including feature selection, grid search in hyperparameters and case balancing could be used to further improve the machine learning algorithm (LeDell et al., 2019b; Torgo, 2010b). In the current study, all areas under the ROC curves were greater than 0.70, most of which were above 0.75. Although there is no gold standard to define a “good” predictive performance, generally, an AUC > 0.70 for a predictive model can be deemed appropriate. Our AUC estimates are within the range of those for predictive models reported in existing studies (Li et al., 2022; Saatchi et al., 2022; Song et al., 2021; Tang et al., 2022) which usually are in clinical settings. Compared with clinical predictors which generally are used for individual-based outcome predictions, environmental predictors are of a greater public health significance which are usually predictive for outcomes at the population level where clinical variables may not be feasible. In addition, consistently appropriate AUCs across different data splits suggested that our models were consistently effective in both internal and external settings.

Strength and uncertainties

This study has several strengths including the assessment of multiple important health outcomes on millions of hospital admissions across the entire New York State and, to the best of our knowledge, the first study investigating the dominant neighborhood environmental characteristics shaping the population vulnerability to adverse health impact

during power outages. The diverse racial/ethnic background of the population and socioeconomic conditions across the state increases the generalizability of our findings. In regards to the methodology, we used sophisticated machine learning algorithms to overcome limitations such as collinearity and nonlinearity issues among the traditional models, and applied multiple strategies including grid search and case balancing to boost the performance of the predictive models. This study provides a standard framework for the application of machine learning methods to environmental health studies.

Although our study provides new insights, some uncertainties should be acknowledged. First, the negative health impacts of power outages may be confounded by reduced air pollution concentrations as select power plants may shut down during some outages (Marufu et al., 2004) and by effects from the extreme weather events which were the most common causes of power outages. While this issue usually was not considered in previous research, we minimized the confounding of air pollution and weather events by integrating these additional pieces of information from different sources and controlling them in the assessment of the health risk of power outages. Second, this study was limited to evaluating the impacts on hospital admissions of cardiovascular, respiratory and intestinal illness and injury, which did not include any clinic or emergency department visits. Thus, this study may be missing some percentage of health impacts from power outages. However, as severely injured residents and those with severe exacerbation of cardiovascular, respiratory or intestinal symptoms generally require immediate care and treatment through hospital admission, our study captures the most severe groups. Third, the statistical significance of *RR* estimates was not considered in the definition of risk elevated communities since the statistical significance was largely subject to the sample size. However, the small sample size at the fine division level should be acknowledged. Future work should confirm these findings by taking the statistical significance of effect estimates into consideration.

Conclusion

Based on the optimal models, we found greater adverse health impacts of power outages in areas with a higher degree of urbanization. The trend was consistent across different health outcomes, with specifically the percent of residents having access to the public transportation, population density, and the percent of minority and soil characteristics being dominant predictors. The optimal predictive models indicated a higher possibility of health risk elevation during power outages for more urbanized communities in downstate counties and those located in the northwest of the state.

Author contributions

S.L., S.S. and Y.H. designed the study. W.Z. and X.D. did the data analyses and wrote the first draft of the report. J.B., H.C., E.S., X.R., Z.G., G.D., K.Z. and R.R. revised the report for important intellectual content.

Ethical approval

All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards.

Conflict of interest

All authors declare they have no actual or potential competing financial interest.

Acknowledgements

This work was supported by Grant # 1R15ES02800001A1 from the National Institute of Environmental Health Sciences of United States. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. We thank the New York State Department of Health for providing the comprehensive health data (data sharing protocol number: 1509-01 A), and thank the New York Department of Public Service for providing the statewide power outage data.

References

Anderson, G.B., Bell, M.L., 2012a. Lights out: impact of the August 2003 power outage on mortality in New York, NY. Epidemiology 23, 189–193. doi:10.1097/EDE.0b013e318245c61c.

Anderson, G.B., Bell, M.L., 2012b. Lights out: impact of the August 2003 power outage on mortality in New York, NY. Epidemiology 23, 189–193. doi:10.1097/EDE.0b013e318245c61c.

Anderson, G.B., Dominici, F., Wang, Y., McCormack, M.C., Bell, M.L., Peng, R.D., 2013. Heat-related emergency hospitalizations for respiratory diseases in the Medicare population. Am. J. Respir. Crit. Care Med. 187, 1098–1103. doi:10.1164/rccm.201211-1969OC.

Barry E, F., Edward W, G., ElaineJ, H., Janet L, H., Brian, L., 2011a. A social vulnerability index for disaster management. J. Homel. Secur. Emerg. Manag. doi:10.2202/1547-7355.1792.

Barry E, F., Edward W, G., ElaineJ, H., Janet L, H., Brian, L., 2011b. A social vulnerability index for disaster management. J. Homel. Secur. Emerg. Manag. doi:10.2202/1547-7355.1792.

Barry E, F., Edward W, G., ElaineJ, H., Janet L, H., Brian, L., Flanagan, B.E., Gregory, E.W., Hallisey, E.J., Heitgerd, J.L., Lewis, B., 2011c. A social vulnerability index for disaster management. J. Homel. Secur. Emerg. Manag. 8. doi:10.2202/1547-7355.1792.

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L., Armstrong, B., 2013a. Time series regression studies in environmental epidemiology. Int. J. Epidemiol. 42, 1187–1195. doi:10.1093/ije/dyt092.

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L., Armstrong, B., 2013b. Time series regression studies in environmental epidemiology. Int. J. Epidemiol. 42, 1187–1195. doi:10.1093/ije/dyt092.

Bloom, M.S., Palumbo, J., Saiyed, N., Lauper, U., Lin, S., 2016. Food and waterborne disease in the Greater New York City area following hurricane Sandy in 2012. Disaster Med. Public Health Prep. 10, 503–511. doi:10.1017/dmp.2016.85.

Bocinsky, R.K., 2019. FedData: Functions to Automate Downloading Geospatial Data Available from Several Federated Data Sources.

Bolin, B., Kurtz, L.C., 2018. Race, Class, Ethnicity, and Disaster Vulnerability. Springer, Cham, pp. 181–203. doi:10.1007/978-3-319-63254-4_10.

Chiras, D., 2009. Power from the Wind: Achieving Energy Independence. New Society Publishers.

Christine, D., Kathryn, L., Sarah, J., Kazuhiko, I., Thomas, M., 2019a. Health impacts of citywide and localized power outages in New York City. Environ. Health Perspect. 126, 67003. doi:10.1289/EHP2154.

Christine, D., Kathryn, L., Sarah, J., Kazuhiko, I., Thomas, M., 2019b. Health impacts of citywide and localized power outages in New York City. Environ. Health Perspect. 126, 67003. doi:10.1289/EHP2154.

Cutter, S.L., Ash, K.D., Emrich, C.T., 2016. Urban–rural differences in disaster resilience. Ann. Am. Assoc. Geogr. 106, 1236–1252. doi:10.1080/24694452.2016.1194740.

Di, Q., Wang, Yan, Zanobetti, A., Wang, Yun, Koutrakis, P., Choirat, C., Dominici, F., Schwartz, J.D., 2017. Air pollution and mortality in the medicare population. N. Engl. J. Med. 376, 2513–2522. doi:10.1056/NEJMoa1702747.

Elith, J., Leathwick, J.R., Hastie, T., 2008. A working guide to boosted regression trees. J. Anim. Ecol 77, 802–813. doi:10.1111/j.1365-2656.2008.01390.x.

Fang, J., Su, Y., 2019. Effects of soils and irrigation volume on maize yield, irrigation water productivity, and nitrogen uptake. Sci. Rep. 9, 7740. doi:10.1038/s41598-019-41447-z.

Gasparrini, A., 2013. Distributed Lag Linear and Non-Linear Models for Time Series Data Document Is Available at R Project.

Gasparrinia, A., Armstrong, B., Kenward, M.G., 2010. Distributed lag non-linear models. Stat. Med. 29, 2224–2234. doi:10.1002/sim.3940.

GISP, C.T.E.P.D., n.d. Social Vulnerability and Hazard Analysis for Hurricane Harvey. University of Central Florida.

Hagyó, A., Tóth, G., 2018. The impact of environmental policy on soil quality: Organic carbon and phosphorus levels in croplands and grasslands of the European Natura 2000 network. J. Environ. Manage. 223, 9–15. doi:10.1016/j.jenvman.2018.06.003.

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., Coulston, J., Herold, N., Wickham, J., Megown, K., 2015. Completion of the 2011 National Land Cover Database for the Conterminous United States - representing a decade of land cover change information. Photogramm. Eng. Remote Sens. doi:10.14358/PERS.81.5.345.

Kaur, R.R., Luthra, A., 2018. Population growth, urbanization and electricity - Challenges and initiatives in the state of Punjab, India. Energy Strategy Reviews 21, 50–61. doi:10.1016/j.esr.2018.04.005.

Kishore, N., Marqués, D., Mahmud, A., Kiang, M.V, Rodriguez, I., Fuller, A., Ebner, P., Sorensen, C., Racy, F., Lemery, J., Maas, L., Leaning, J., Irizarry, R.A., Balsari, S., Buckee, C.O., 2018a. Mortality in Puerto Rico after Hurricane Maria. N. Engl. J. Med. 379, 162–170. doi:10.1056/NEJMsa1803972.

Kishore, N., Marqués, D., Mahmud, A., Kiang, M.V, Rodriguez, I., Fuller, A., Ebner, P., Sorensen, C., Racy, F., Lemery, J., Maas, L., Leaning, J., Irizarry, R.A., Balsari, S., Buckee, C.O., 2018b. Mortality in Puerto Rico after Hurricane Maria. N. Engl. J. Med. 379, 162–170. doi:10.1056/NEJMsa1803972.

LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., Malohlava, M., 2019a. H_2O: R Interface for "H_2O.".

LeDell, E, Gill, N, Aiello, S, Fu, A, Candel, A, Click, C, Kraljevic, T, Nykodym, T, Aboyoun, P, Kurka, M, Malohlava, M, 2019b. H_2O: R Interface for "H_2O.".

Li, L., Tian, H., Zhang, B., Wang, W., Li, B., 2022. Prediction for distant metastasis of breast cancer using dynamic contrast-enhanced magnetic resonance imaging images under deep learning. Comput. Intell. Neurosci. 2022, 6126061. doi:10.1155/2022/6126061.

Li, M., Shaw, B.A., Zhang, W., Vasquez, E., Lin, S., 2019. Impact of Extremely Hot Days on Emergency Department Visits for Cardiovascular Disease among Older Adults in New York State. Int. J. Environ. Res. Public Health 16. doi:10.3390/ijerph16122119.

Lin, S., Fletcher, B.A., Luo, M., Chinery, R., Hwang, S.-A.A., 2011a. Health impact in New York City during the Northeastern blackout of 2003. Public Health Reports 126, 384–393. doi:10.1177/003335491112600312.

Lin, S., Fletcher, B.A., Luo, M., Chinery, R., Hwang, S.-A.A., 2011b. Health impact in New York City during the Northeastern blackout of 2003. Public Health Rep. 126, 384–393. doi:10.1177/003335491112600312.

Lobdell, D.T., Jagai, J.S., Rappazzo, K., Messer, L.C., 2011. Data sources for an environmental quality index: availability, quality, and utility. Am. J. Public Health 101, S277–S285. doi:10.2105/AJPH.2011.300184.

Logan, J.R., Xu, Z., 2015. Vulnerability to hurricane damage on the U.S. Gulf coast since 1950. Geograph. Rev. 105, 133–155. doi:10.1111/j.1931-0846.2014.12064.x.

Marufu, L.T., Taubman, B.F., Bloomer, B., Piety, C.A., Doddridge, B.G., Stehr, J.W., Dickerson, R.R., 2004. The 2003 North American electrical blackout: An accidental experiment in atmospheric chemistry. Geophys. Res. Lett. 31. doi:10.1029/2004GL019771.

Muncan, B., 2018. Cardiovascular disease in racial/ethnic minority populations: illness burden and overview of community-based interventions. Public Health Rev. 39, 32. doi:10.1186/s40985-018-0109-4.

Nayak, A., Islam, S.J., Mehta, A., Ko, Y.-A., Patel, S.A., Goyal, A., Sullivan, S., Lewis, T.T., Vaccarino, V., Morris, A.A., Quyyumi, A.A., 2020. Impact of social vulnerability on COVID-19 incidence and outcomes in the United States. medRxiv doi:10.1101/2020.04.10.20060962.

Nayak, S.G., Shrestha, S., Kinney, P.L., Ross, Z., Sheridan, S.C., Pantea, C.I., Hsu, W.H., Muscatiello, N., Hwang, S.A., 2018. Development of a heat vulnerability index for New York State. Public Health 161, 127–137. doi:10.1016/j.puhe.2017.09.006.

New York–New Jersey Information Office, 2016. County Employment and Wages in New York — Second Quarter 2016 [WWW Document]. U.S. Bureau of Labor Statistics.

Nishar, A., Bader, M.K.-F., O'Gorman, E.J., Deng, J., Breen, B., Leuzinger, S., 2017. Temperature effects on biomass and regeneration of vegetation in a geothermal area. Front. Plant Sci. 8, 249. doi:10.3389/fpls.2017.00249.

Parker, J.D., Kravets, N., Vaidyanathan, A., 2018. Particulate matter air pollution exposure and heart disease mortality risks by race and ethnicity in the United States: 1997 to 2009 National Health Interview Survey with mortality follow-up through 2011. Circulation 137, 1688–1697. doi:10.1161/CIRCULATIONAHA.117.029376.

Rich, D.Q., Zhang, W., Lin, S., Squizzato, S., Thurston, S.W., van Wijngaarden, E., Croft, D., Masiol, M., Hopke, P.K., 2019. Triggering of cardiovascular hospital admissions by source specific fine particle concentrations in urban centers of New York State. Environ. Int. 126, 387–394. doi:10.1016/J.ENVINT.2019.02.018.

Román, M.O., Stokes, E.C., Shrestha, R., Wang, Z., Schultz, L., Carlo, E.A.S., Sun, Q., Bell, J., Molthan, A., Kalb, V., Ji, C., Seto, K.C., McClain, S.N., Enenkel, M., 2019. Satellite-based assessment of electricity restoration efforts in Puerto Rico after Hurricane Maria. PLoS One 14, e0218883.

Saatchi, M., Khatami, F., Mashhadi, R., Mirzaei, A., Zareian, L., Ahadi, Z., Aghamir, S.M.K., 2022. Diagnostic accuracy of predictive models in prostate cancer: a systematic review and meta-analysis. Prostate Cancer 2022, 1742789. doi:10.1155/2022/1742789.

Sheridan, S.C., Zhang, W., Deng, X., Lin, S., 2021. The individual and synergistic impacts of windstorms and power outages on injury ED visits in New York State. Sci. Total Environ. 797, 149199. doi:10.1016/j.scitotenv.2021.149199.

Sidey-Gibbons, J.A.M., Sidey-Gibbons, C.J., 2019. Machine learning in medicine: a practical introduction. BMC Med. Res. Method. 19, 64. doi:10.1186/s12874-019-0681-4.

Song, X., Liu, X., Liu, F., Wang, C., 2021. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. Int. J. Med. Inform. 151, 104484. doi:10.1016/j.ijmedinf.2021.104484.

Tang, M., Gao, L., He, B., Yang, Y., 2022. Machine learning-based prognostic prediction models of non-metastatic colon cancer: analyses based on surveillance, epidemiology and end results database and a chinese cohort. Cancer Manag. Res. 14, 25–35. doi:10.2147/CMAR.S340739.

The NYS GIS Program Office, 2017. The Street and Address Maintenance (SAM) Program [WWW Document].

Tian, X., Li, X., Huang, Y., Guo, P., Li, M., Zhang, W., Du, Z., Chong, Y., Hao, Y., 2019. Using machine learning algorithms to predict Hepatitis B surface antigen seroclearance. Comput. Math. Methods Med.

Torgo, L., 2010a. Data Mining with R, Learning with Case Studies. Chapman and Hall/CRC.

Torgo, L, 2010b. Data Mining with R, Learning with Case Studies. Chapman and Hall/CRC.

Wikipedia contributors, 2019. List of United States Counties by per Capita Income.

Xiao, X., Gasparrini, A., Huang, J., Liao, Q., Liu, F., Yin, F., Yu, H., Li, X., 2017. The exposure-response relationship between temperature and childhood hand, foot and mouth disease: A multicity study from mainland China. Environ. Int. 100, 102–109. doi:10.1016/J.ENVINT.2016.11.021.

Zhang, W., Du, Z., Zhang, D., Yu, S., Hao, Y., 2016a. Boosted regression tree model-based assessment of the impacts of meteorological drivers of hand, foot and mouth disease in Guangdong. Sci. Total Environ. 553. doi:10.1016/j.scitotenv.2016.02.023.

Zhang, W., Du, Z., Zhang, D., Yu, S., Huang, Y., Hao, Y., 2016b. Assessing the impact of humidex on HFMD in Guangdong Province and its variability across social-economic status and age groups. Sci. Rep. 6. doi:10.1038/srep18965.

Zhang, W., Kinney, P.L., Rich, D.Q., Sheridan, S.C., Romeiko, X.X., Dong, G., Stern, E.K., Du, Z., Xiao, J., Lawrence, W.R., Lin, Z., Hao, Y., Lin, S., 2020a. How community vulnerability factors jointly affect multiple health outcomes after catastrophic storms. Environ. Int. 134. doi:10.1016/j.envint.2019.105285.

Zhang, W., Lin, S., Hopke, P.K., Thurston, S.W., van Wijngaarden, E., Croft, D., Squizzato, S., Masiol, M., Rich, D.Q., 2018. Triggering of cardiovascular hospital admissions by fine particle concentrations in New York state: Before, during, and after implementation of multiple environmental policies and a recession. Environ. Pollut. 242, 1404–1416. doi:10.1016/j.envpol.2018.08.030.

Zhang, W., Sheridan, S.C., Birkhead, G.S., Croft, D.P., Brotzge, J.A., Justino, J.G., Stuart, N.A., Du, Z., Romeiko, X.X., Ye, B., Dong, G., Hao, Y., Lin, S., 2020b. Power outage: an ignored risk factor for COPD exacerbations. Chest 158, 2346–2357. doi:10.1016/j.chest.2020.05.555.

Zhang, W., Sheridan, S.C., Birkhead, G.S., Croft, D.P., Brotzge, J.A., Justino, J.G., Stuart, N.A., Du, Z., Romeiko, X.X., Ye, B., Dong, G., Hao, Y., Lin, S., 2020c. Power outage: an ignored risk factor for COPD exacerbations. Chest 158, 2346–2357. doi:10.1016/j.chest.2020.05.555.

Zhao, P., Zhang, M., 2018. The impact of urbanisation on energy consumption: a 30-year review in China. Urban Clim. 24, 940–953. doi:10.1016/j.uclim.2017.11.005.

Zoraster, R.M., 2010. Vulnerable populations: hurricane Katrina as a case study. Prehospital. Disaster Med. 25, 74–78. doi:10.1017/S1049023X00007718.