ENVIRONMENTAL RESEARCH
HEALTH

PURPOSE-LED PUBLISHING™

Part of the ENVIRONMENTAL RESEARCH SERIES

PAPER • **OPEN ACCESS**

Integrating machine learning-based variable selection into heat vulnerability index design

To cite this article: Shuyue Qu *et al* 2026 *Environ. Res.: Health* **4** 035023

IOP Publishing Environ. Res.: Health 4 (2026) 035023 https://doi.org/10.1088/2752-5309/ae9633

ENVIRONMENTAL RESEARCH
HEALTH

OPEN ACCESS

RECEIVED
22 March 2026

REVISED
13 July 2026

ACCEPTED FOR PUBLICATION
6 August 2026

PUBLISHED
15 September 2026

PAPER

Integrating machine learning-based variable selection into heat vulnerability index design

Shuyue Qu[1,2,*], Jana Sillmann[1,3], Benjamin W Barrett[2,4], Peter M Graffy[2,4], Benjamin Poschlod[1], Lukas Brunner[1], Raed Mansour[2,5], Malte von Szombathely[1], Finley Hay-Chapman[2,6], Teresa H Horton[2,7], Jennifer Chan[2,4], Sheetal Khedkar Rao[2,8], Kyra Woods[2,9], Abel N Kho[2,4] and Daniel E Horton[2,6]

1 Research Unit Sustainability and Climate Risk, Earth and Society Research Hub (ESRAH), University of Hamburg, Hamburg, Germany
2 Defusing Disasters Working Group, Roberta Buffett Institute for Global Affairs, Northwestern University, Evanston, IL, United States of America
3 Center of International Climate Research (CICERO), Oslo, Norway
4 Feinberg School of Medicine, Northwestern University, Chicago, IL, United States of America
5 Metropolitan Planning Council, Chicago, IL, United States of America
6 Department of Earth, Environmental, and Planetary Sciences, Northwestern University, Evanston, IL, United States of America
7 Department of Anthropology, Northwestern University, Evanston, IL, United States of America
8 Department of Medicine, University of Illinois, Chicago, IL, United States of America
9 Sustain Our Future Foundation, Chicago, IL, United States of America
* Author to whom any correspondence should be addressed.

E-mail: shuyue.qu@uni-hamburg.de and shuyue.qu@northwestern.edu

Keywords: heat vulnerability, machine learning, urban heat, urban health, health risk

Supplementary material for this article is available online

Abstract

As climate change intensifies, health risks from extreme heat are rising. Accurate assessment of heat vulnerability at high spatial resolution is crucial for developing effective adaptation strategies, particularly in socioeconomically heterogeneous urban settings. However, the identification of key indicators underlying heat vulnerability remains challenging. Using Chicago, Illinois (USA) as a case study, we systematically compare different variable selection strategies in community-level heat vulnerability assessments. We take the conventional unsupervised principal component analysis-based heat vulnerability index (HVI) as a baseline, and compare it with supervised approaches that incorporate variable selection, including machine learning algorithms (Lasso regression, random forest (RF), and XGBoost) as well as traditional statistical methods (simple linear regression and polynomial regression). Using the vulnerability indicator subsets identified by each variable selection method, we construct multiple HVIs and evaluate their performance against heat-related excess mortality. Our work indicates that supervised variable selection improves the performance of HVIs in capturing heat-related health risks. Among all methods, the RF-based variable selection algorithm achieves the best overall results, highlighting the potential of machine learning to enhance heat vulnerability assessment tools. Our results demonstrate that poverty rate, lack of air conditioning, and proportion of residents aged 65 and above consistently emerged as important determinants of heat vulnerability in Chicago.

1. Introduction

In recent years, climate change has exacerbated the frequency, intensity, and duration of extreme heat events (IPCC 2023, World Meteorological Organization 2023, World Health Organization 2024). Substantial epidemiological evidence suggests that heat is closely linked to increased risks of multiple adverse health outcomes, including cardiovascular and respiratory diseases, renal disorders, adverse pregnancy outcomes, and mental health conditions such as anxiety and depression (Benmarhnia *et al* 2015, Bell *et al* 2024). Heat-related mortality has been identified as one of the most serious climate threats to human health (Lüthi *et al* 2023), and with continued global warming, heat-related morbidity and mortality are projected to rise (Ebi *et al* 2021). Given current and impending climate and health challenges, accurate assessments of heat-related health risks are critical for guiding targeted and effective public health interventions and adaptations.

However, not all populations are at equal health risk from heat (Reid *et al* 2009). Previous studies indicate that heat vulnerability is associated with a wide range of factors, including socioeconomic conditions, demographic characteristics, and the built environment (Reid *et al* 2009, Chuang and Gober 2015, Niu *et al* 2021, Graffy *et al* 2024). Because direct measures of heat vulnerability are unavailable, researchers have developed proxy-based tools, such as heat vulnerability indices (HVIs), that integrate multidimensional indicators of heat-related health risk to identify the populations and communities most vulnerable to heat. HVIs are increasingly used to guide where localities prioritize investments such as cooling access, outreach, housing improvements, and greening initiatives.

One of the most influential HVI frameworks was proposed by Reid *et al* (2009), who developed a principal component analysis (PCA)-based HVI at the census tract level for the United States (U.S.). This unsupervised approach applies PCA to pre-selected heat vulnerability indicators to identify latent factors that explain the largest variance, without reference to health outcome data. These factors are then interpreted as representing underlying dimensions in heat vulnerability and aggregated to obtain cumulative HVI scores (Reid *et al* 2009, Tate 2012, Cheng *et al* 2021). The Reid *et al* (2009) methodology has since been widely adopted in heat vulnerability assessments, both in the U.S. and in other global settings (Conlon *et al* 2020, Manware *et al* 2022, Wang *et al* 2023). However, this methodology has certain limitations, including reliance on pre-selected and subjective indicators and lack of validation against health outcomes (Tate 2012, Wolf *et al* 2014, Cheng *et al* 2021, Niu *et al* 2021), thereby raising concerns about whether unsupervised HVIs adequately capture heat-related health risks.

Building on the unsupervised HVI framework proposed by Reid *et al* (2009), Conlon *et al* (2020) introduced a supervised HVI that incorporates health outcome data in the index construction. This approach linearly regresses each potential vulnerability indicator against the proportion of mortality on extreme heat days and applies the PCA to only those indicators that were moderately significantly associated ($p < 0.20$) with mortality. However, both the supervised and unsupervised HVIs explained only a small proportion of the variation in mortality on extreme heat days in Detroit, Michigan (Conlon *et al* 2020). This limited performance may indicate that variable selection based on simple linear regression (SLR) cannot fully capture the relationships between heat vulnerability indicators and heat-related health outcomes, suggesting the need to explore nonlinear associations and other modeling approaches. Indeed, a recent study demonstrated that nonlinear models more accurately characterize the relationships between socioeconomic variables and health outcomes compared to linear models (Jeong *et al* 2024), which motivates the use of more advanced and flexible variable selection strategies.

In this context, machine learning offers a promising methodological direction for environmental health research and risk assessment (Potash *et al* 2015, 2020, Sadilek *et al* 2018, Graffy *et al* 2026). While not inherently superior to traditional statistical methods, machine learning approaches can provide locally validated, outcome-informed screening of indicators. Beyond the limitations of SLR, machine learning-based variable selection methods are capable of identifying complex relationships among variables, evaluating variable importance relative to health outcomes during model training, and balancing computational efficiency with overfitting resistance (Guo and Zhou 2022, Zhao *et al* 2023, Theng and Bhoyar 2024, Zhou *et al* 2024). These capabilities suggest that machine learning-based variable selection methods may help refine the indicators used in HVI construction.

This study incorporates machine learning-based variable selection methods into the conventional PCA-based HVI framework to explore the effectiveness of different variable selection strategies. Within the PCA-based HVI framework, three approaches are evaluated and compared: (a) unsupervised HVI using pre-selected indicators from Reid *et al* (2009); (b) supervised HVI that uses traditional statistical methods for variable selection, including SLR (following Conlon *et al* 2020) and polynomial regression (PR); and (c) supervised HVI that uses machine learning-based methods for variable selection, including Least Absolute Shrinkage and Selection Operator (Lasso) regression, random forest (RF), and XGBoost. We apply these approaches to the 77 community areas (CAs) in Chicago, Illinois, which serves as an ideal study area due to the availability of high-resolution environmental, socioeconomic, demographic, and health data at the CA level. The performance of the HVIs derived from these different approaches is systematically validated against CA-level heat-related excess mortality, providing methodological insights and recommendations for future heat vulnerability research. Given that HVI maps influence the prioritization of community resources, this study emphasizes locally justifiable indicators and outcome-informed evaluation, both of which have important equity implications.

2. Data

The data for this study are collected and processed at the Chicago CA level using the CA boundaries obtained from the Chicago Data Portal (shown in supplemental figure S1). CAs are stable geographical divisions in Chicago that have been largely unchanged since the 1920s. Many demographic and socioeconomic indicators in Chicago are analyzed and released at this level, providing comparable data across the 77 different CAs over time. Based on these CA boundaries, we integrated health, demographic, and socioeconomic datasets, and aggregated meteorological and land use data to the CA level for subsequent analyses.

2.1. Meteorological data

Daymet Version 4 R1 contains North American daily meteorological data from 1980 to present at a resolution of 1 km. This dataset provides gridded weather estimates using statistical modeling techniques to interpolate and extrapolate ground-based observations (Thornton *et al* 2022). It is a research product of the Environmental Sciences Division at Oak Ridge National Laboratory, and widely used in ecosystem modeling and climate change research (Walton and Hall 2018). We extracted daily minimum temperature (T_{min}), maximum temperature (T_{max}), and vapor pressure (V_{p}) data for the hot season (May to September) from 1993 to 2019. These daily gridded estimates were then spatially aggregated to the CA level by unweighted averaging of grid cells whose centroids fell within each CA polygon in Chicago.

2.2. Mortality data

Mortality data used in this study contain individual-level records on age, sex, race/ethnicity, and CA of residence for decedents in Chicago from 1993 to 2019. These data were provided by the Illinois and Chicago Departments of Public Health (IDPH and CDPH). The use of mortality data was governed by Data Use Agreements with IDPH and CDPH.

2.3. Population data

Population data used in this study were obtained from the standardized time series tables provided by the National Historical Geographic Information System, a project of the Integrated Public Use Microdata Series at the University of Minnesota (Manson *et al* 2024). These datasets geo-standardize the 1990, 2000, and 2010 decennial census data to the 2010 census tract boundaries, providing fine-grained population information over time. Annual population estimates were generated using linear interpolation, assuming linear population change between adjacent decennial census years. For the years after 2010 (2011–2019), we used the American Community Survey (ACS) five-year estimates as supplements. All population data were aggregated to the Chicago CA boundaries to ensure geographic consistency. Chicago population characteristics and all-cause mortality counts for the study period are presented in table 1.

2.4. Heat vulnerability indicators

Heat vulnerability indicators refer to measurable socioeconomic, demographic, and environmental variables that represent specific dimensions of community-level heat vulnerability and serve as the candidate inputs for constructing the composite HVI. We adapted the ten heat vulnerability indicators proposed by Reid *et al* (2009). These indicators were organized into four categories: diabetes prevalence (*Diabetes*), demographic/socioeconomic characteristics (*Race other than White*, *Age > 65 years*, *Live alone*, *Age > 65 living alone*, *Below poverty line*, *Less than high school diploma*), land cover (*Not Green Space*), and air conditioning (AC) access (*No central AC*, *No AC of any kind*). Due to differences in data availability and measurement across datasets, the operationalization of several indicators differs slightly between the original (Reid *et al* 2009) framework and our work (detailed crosswalk is provided in supplementary table S1).

Specifically, the diabetes indicator was represented by *Diabetes Prevalence*, obtained from the annual Healthy Chicago Survey conducted by the CDPH. This indicator is defined as the percent of adults who reported that they had ever been diagnosed with diabetes by a doctor, nurse, or other health professional, excluding pre-diabetes and diabetes only during pregnancy.

Data on age, race/ethnicity, and educational attainment were obtained from the Community Data Snapshots provided by the Chicago Metropolitan Agency for Planning. To better reflect racial and ethnic disparities in heat vulnerability, we split the '*Race other than White*' indicator into two specific indicators: the proportion of Non-Hispanic Black residents (*Black (Non-Hispanic)*) and the proportion of Hispanic or Latino residents (*Hispanic or Latino (of Any Race)*). Proportions of the population aged over 65 years (*Age above 65*), individuals living alone (*Living Alone*), older adults living alone (*Age above 65 and Living*

Table 1. Population characteristics in Chicago for census years 1990, 2000, 2010, and 2020, and total May to September mortality counts for 1993–2019. In the years 2000, 2010, and 2020, some individuals were categorized as belonging to multiple racial groups. To accurately represent the distribution across racial groups, these individuals were counted in each of their respective racial categories; therefore, the sum across racial groups exceeds the total population count. For the mortality data, some individual-level characteristics are missing for a small number of records, and therefore category totals may not necessarily equal the total mortality count.

	Population (1990)	Population (2000)	Population (2010)	Population (2020)	All-Cause Mortality (1993–2019)
Total	2783 726	2896 016	2695 598	2746 388	251 706
Female	1449 021	1490 909	1387 526	1413 663	120 535
Male	1334 705	1405 107	1308 072	1332 725	131 167
Age 65+	330 182	298 803	277 932	350 550	156 719
Age 0–64	2453 544	2597 213	2417 666	2395 838	94 949
Hispanic	535 315	753 644	756 350	819 518	22 887
Non-Hispanic	2248 411	2142 372	1939 248	1926 870	224 370
Black	1087 711	1084 221	913 009	845 639	114 659
White	1263 524	1282 320	1270 097	1253 782	129 524
Asian or Pacific Islander	104 118	145 132	170 540	226 632	5 843
American Indian or Alaskan Native	7 064	20 898	26 933	71 573	282

Alone), and educational level (*Less than High School Diploma*) were defined in alignment with Reid *et al* (2009). *Poverty Rate* was derived from the CDPH Chicago Health Atlas platform and is defined as the proportion of residents living in households below the federal poverty threshold. These indicators were based on the ACS five-year census tract-level estimates and are reported as percentages at the CA level.

For the vegetation indicator *Not Green Space*, we used land cover data from the National Land Cover Database (NLCD) covering the Chicago area for the years 1993–2022. The NLCD is a land cover dataset derived from Landsat satellite imagery, with a spatial resolution of 30 m. We aggregated the land cover data for each year at the CA level. Each pixel was assigned to the CA polygon in which its centroid was located. Following the classification used by Reid *et al* (2009), green space is defined based on selected land cover classes: deciduous forest, evergreen forest, mixed forest, dwarf scrub, orchards/vineyards/other, pasture/hay, small grains, fallow, row crops, urban/recreational grasses, palustrine forested wetlands, and palustrine scrub/shrub wetlands. The percentage of green space for each CA was calculated as the total area of these land cover classes divided by the total CA area. The percentage of *Not Green Space* was then calculated as 100 minus the percentage of green space. Finally, a multi-year average of *Not Green Space* was computed for each CA over the period 1993–2022 to represent long-term non-green space coverage.

The AC access indicator (*No AC Access*) was obtained from the Healthy Chicago Survey, which for the first time in 2023 included a question on household AC availability: 'Is any AC equipment used in your home?'. The questionnaire results provide the proportion of 'yes' and 'no' responses. To ensure consistent variable directionality, *No AC Access* was calculated as 100 minus the percentage of AC access within each CA. Given the lack of available data on central AC use at the CA level, and because the survey question captures all forms of AC, *No AC Access* was adopted as the only indicator of overall AC access in this study.

In summary, 10 heat vulnerability indicators were included in our analysis (table 2). All indicators were averaged over their available data periods. Correlations among these heat vulnerability indicators are presented in supplemental figure S2.

3. Methods

3.1. Identification of heat days

Previous studies have shown that the most severe health impacts of heatwaves often happen during nighttime heat exposure (Sarofim *et al* 2016). A threshold of 70 °F (21.1 °C) was adopted to define warm nights in this study. This threshold is consistent with those used in climate assessments such as NCA5 (U.S. Global Change Research Program 2023). We used the daily minimum heat index (HI_{min}) instead of air temperature as the exposure metric. The HI combines air temperature (T) and relative humidity (RH) and can better characterize the 'feels like' temperature of the human body.

The HI_{min} for each CA was calculated based on the daily minimum air temperature ($T_{\min}$) and RH following the National Weather Service Rothfusz regression equation (NOAA 2022). To determine the RH input, we first computed a weighted daily mean temperature (T_{mean}) as

Table 2. Data period, resources, and ranges for the heat vulnerability indicators used in this study.

Indicator	Data Period	Data Resource	Universe	Percent Mean (Range)
Diabetes Prevalence	2021–2022	Chicago Department of Public Health	Adults	12.2 (0–38.2)
Black (Non-Hispanic)	2017–2021	American Community Survey five-year estimates	Total population	36.7 (0.4–96.9)
Hispanic or Latino (of Any Race)	2017–2021	American Community Survey five-year estimates	Total population	26.7 (0.4–90.9)
Age above 65	2017–2021	American Community Survey five-year estimates	Total population	14.0 (5.3–29.5)
Living Alone	2016–2020	American Community Survey five-year estimates	Total population	14.9 (4.8–35.0)
Age above 65 and Living Alone	2018–2022	American Community Survey five-year estimates	Population older than 65 years	36.9 (11.0–71.4)
Poverty Rate	2018–2022	American Community Survey five-year estimates	Total population	18.6 (3.9–52.1)
Less than High School Diploma	2017–2021	American Community Survey five-year estimates	Population 25 years and older	14.8 (1.7–38.9)
Not Green Space	1993–2022	National Land Cover Database	Total land area	99.1 (87.5–100.0)
No AC Access	2022–2023	Chicago Department of Public Health	Households	12.9 (0.9–33.2)

$0.606 \times T_{\text{max}} + 0.394 \times T_{\text{min}}$ (Spangler *et al* 2019). Saturated vapor pressure (e_s, in Pa) was then derived from T_{mean} using a Magnus–Tetens formula:

$$e_s = 610.78 \times \exp\left(\frac{17.269 \times T_{\text{mean}}}{T_{\text{mean}} + 237.3}\right).$$

RH (%) was calculated as the ratio of actual vapor pressure (V_p) to e_s:

$$\text{RH} = 100 \times \left(\frac{V_p}{e_s}\right).$$

The resulting RH was applied to T_{min} to compute the raw HI_{min}. To capture the acute effect of heat, a two-day moving average was applied to these raw HI_{min} values (Barnett and Åström 2012, Conlon *et al* 2020). A 'heat day' was defined as a day when HI_{min} exceeded 70 °F (21.1 °C) for at least two consecutive days. This approach captures both the immediate physiological impact of short-term heat exposure and the sustained nature of extreme heat events.

3.2. Heat-related excess mortality

We estimated heat-related excess mortality by defining the observed number of deaths (O) as the daily average all-cause mortality during the index period (heat days) and the expected number of deaths (E) as the daily average all-cause mortality during the reference period (non-heat days). Excess mortality was then calculated as $O - E$ (Fouillet *et al* 2006, Azhar *et al* 2014, Stang *et al* 2020). Analyses were restricted to the hot season (May to September) for each year. To avoid interference from the COVID-19 pandemic, the time window for heat-related excess mortality estimation was limited to 1993–2019. Annual average heat-related excess mortality was calculated by multiplying the daily values by the total number of heat days in the corresponding year. Annual and daily time series of heat-related excess mortality were generated to examine temporal trends. Heat-related excess mortality was mapped at the CA level, and Moran's I was calculated to evaluate spatial autocorrelation.

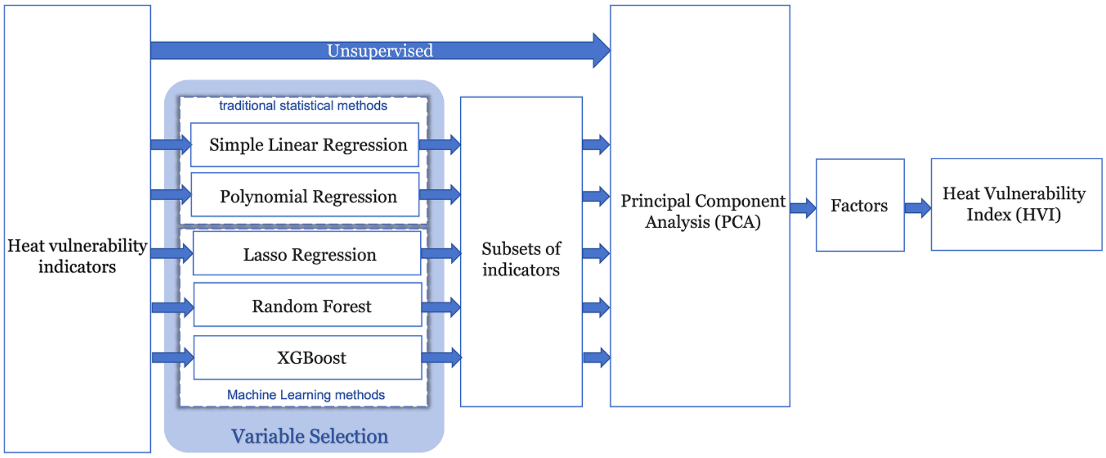

Figure 1. Workflow of supervised and unsupervised heat vulnerability index (HVI) construction. In the unsupervised framework, the complete set of candidate heat vulnerability indicators is directly fed into principal component analysis (PCA) to extract latent factors and further construct the HVI. In the supervised framework, heat-related excess mortality is incorporated as the outcome variable, and the candidate heat vulnerability indicators first undergo a data-driven variable selection step. Five supervised variable selection methods (simple linear regression, polynomial regression, Lasso regression, random forest, and XGBoost) are used to select subsets of indicators associated with heat-related excess mortality. PCA is then separately applied to each indicator subset to derive factors for HVI construction.

3.3. Unsupervised heat vulnerability index

The PCA-based HVI framework proposed by Reid *et al* (2009) is an unsupervised approach that generates HVI scores from a set of pre-selected heat vulnerability indicators. PCA is a dimensionality reduction method that transforms multiple potentially correlated variables into a smaller set of uncorrelated components, capturing most of the variance in the original data (Sun *et al* 2017, Mallen *et al* 2019, Li *et al* 2025). It is a popular approach in the construction of many societal vulnerability indices (Tate 2012).

In this framework, we performed the PCA on the heat vulnerability indicators using Python's Scikit-learn PCA tool. All indicators were standardized using z-scores prior to PCA to ensure comparability across variables with different scales. Varimax rotation was used to improve the interpretability of the retained principal components while preserving the orthogonality among factors. Following the approach of Reid *et al* (2009), the number of principal components retained was based on a combination of eigenvalues greater than one, significant breaks in the scree plot, and a cumulative explained variance exceeding 75%. We refer to these retained principal components as 'factors.' Subsequently, each CA was assigned a score from 1 to 6 for each component according to distance from the mean, with scores increasing from the lowest to the highest category of standardized factor scores. These factor scores were equally weighted and summed to calculate the cumulative HVI scores for each CA in Chicago.

3.4. Supervised heat vulnerability indices

For supervised approaches, the same candidate indicators from the unsupervised approach were used, and heat-related excess mortality was added as the response variable. Compared to the unsupervised approach, supervised HVI involves a data-driven variable selection step, where heat vulnerability indicators are the predictors and heat-related excess mortality is the health outcome. This study integrates three machine learning-based variable selection methods (Lasso regression, RF, and XGBoost) into the PCA-based HVI framework, and compares them with the results derived from the unsupervised HVI and HVIs that use more traditional supervised variable selection methods, i.e., SLR and PR (figure 1). For machine learning-based approaches, internal validation procedures (e.g. cross-validation and out-of-bag (OOB) evaluation) were applied where appropriate.

3.4.1. Simple linear regression and polynomial regression

We first fitted univariate SLR models using ordinary least squares (OLS) to identify initial linear associations between each candidate indicator and heat-related excess mortality. Indicators with statistically significant coefficients ($p < 0.05$) were retained. To capture potential nonlinear relationships, we further employed PR, an extension of SLR that incorporates polynomial terms (Hastie *et al* 2009). We used a univariate PR model of degree 3 (including linear, quadratic, and cubic terms) to regress heat-related excess mortality on each indicator. A joint hypothesis test (F-test) was then used to assess the overall statistical significance of the polynomial terms. Indicators with statistically significant joint F-tests ($p < 0.05$) were retained for HVI construction.

3.4.2. Lasso regression

Lasso regression is a type of regularized linear regression that minimizes the sum of squared residuals with L1 regularization, which shrinks some coefficients and sets others to zero, thereby retaining important features (Tibshirani 1996, Lounici 2008, Cai *et al* 2018, Emmert-Streib and Dehmer 2019, Karpinska *et al* 2021). Compared with OLS, Lasso can mitigate the effects of multicollinearity, reduce overfitting, and potentially improve the overall prediction accuracy and interpretability (Tibshirani 1996, Muthukrishnan and Rohini 2016, Emmert-Streib and Dehmer 2019, Theng and Bhoyar 2024). In this study, the dataset was randomly split into training and testing sets. The mean and standard deviation were calculated from the training set, and these parameters were then used to standardize both the training and testing sets. We employed LassoCV with five-fold cross-validation on the training set to optimize the regularization parameter. Indicators with non-zero coefficients were selected for HVI construction.

3.4.3. Random forest

RF is an ensemble machine learning algorithm that can learn nonlinear relationships and interactions from data without explicit model specification (Grömping 2009, Genuer *et al* 2010, Speiser *et al* 2019). It characterizes complex data structures by constructing multiple decision trees and aggregating their predictions (Breiman 2001). RF provides variable importance measures that reflect the relative contribution of each feature to the fitted model (Nguyen *et al* 2015). In this study, the RF consisted of 5 000 decision trees, and five features were randomly selected at each tree split. The minimum number of split samples was set to five to control the complexity and avoid overfitting. Feature importance was evaluated using the OOB permutation importance, and the top 50% most important indicators were retained for subsequent analysis, which corresponded to an importance threshold of 0.01.

3.4.4. XGBoost

XGBoost is a scalable machine learning method based on gradient boosting decision trees (Chen *et al* 2016). In boosted tree models, gain is a primary metric used to quantify the importance of a feature (Zheng *et al* 2017). The average gain is the total gain of all trees divided by the total number of splits in which the feature is used (Shi *et al* 2019). Higher feature importance scores indicate greater contribution of a feature during model training (Shi *et al* 2019, Jiang *et al* 2023). We employed the built-in, average gain-based feature importance metric of XGBoost to rank the candidate indicators according to their relative contributions to predicting heat-related excess mortality. The dataset was randomly split into training and testing sets. The hyperparameter settings of the model were determined using five-fold cross-validation on the training set. By setting the cumulative importance threshold at 80%, a subset of indicators was selected to retain the most influential indicators while reducing dimensionality. This procedure utilizes XGBoost's capacity to capture nonlinear relationships and feature interactions.

3.5. Evaluation and sensitivity analysis

To validate the resulting HVIs, we first treated both the HVI scores and heat-related excess mortality as continuous variables. Spearman's correlation coefficient was used to test the monotonic association between them. The mean squared error (MSE) and mean absolute error (MAE) were also calculated.

We then converted both the HVI scores and heat-related excess mortality into categorical variables using equal width binning, dividing the variables into low, medium-low, medium-high, and high heat vulnerability categories, with integer scores of 1–4 assigned to these categories accordingly. To assess the ability of the HVIs to differentiate communities across heat vulnerability categories, accuracy and $F1$ scores were employed as classification performance metrics.

To further evaluate the robustness of key heat vulnerability indicators, we conducted sensitivity analyses under different analytical settings: (1) adoption of alternative definitions of heat days, including the use of a daily maximum HI threshold of 110 °F (43.3 °C) and co-occurring daytime and nighttime HI threshold exceedances; and (2) performance of age standardization on heat-related excess mortality based on the 2000 U.S. Standard Population, calculating weighted heat-related deaths by age group to reduce biases from population age structure differences across CAs. These procedures were systematically applied across all HVI approaches, resulting in 30 analytical scenarios (with/without age standardization $\times$ 3 heat day definitions $\times$ 5 variable selection approaches). Across these scenarios, we assessed the consistency of selected indicators to examine the robustness of the selection results.

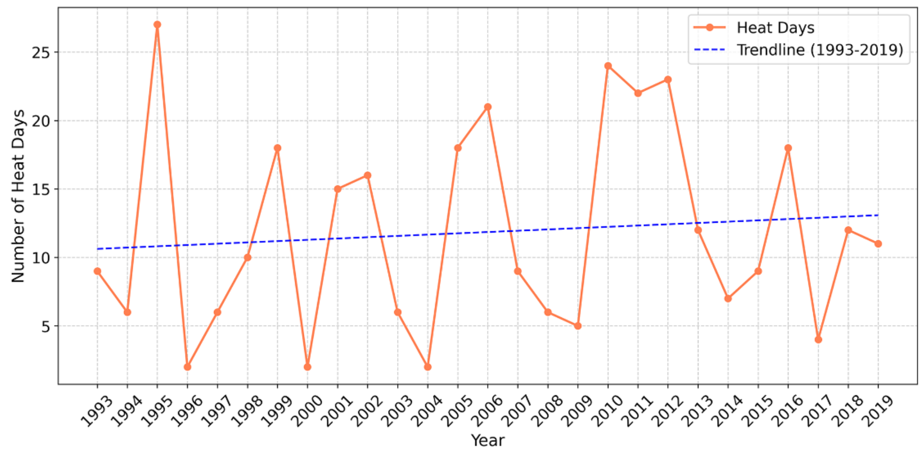

Figure 2. Annual number of heat days and linear trend in Chicago from 1993 to 2019 (slope = 0.09 d yr^{-1}, $p = 0.61$), representing an increase of 0.9 heat days per decade. Heat days are defined as at least two consecutive days with the daily minimum heat index exceeding 70 °F (21.1 °C).

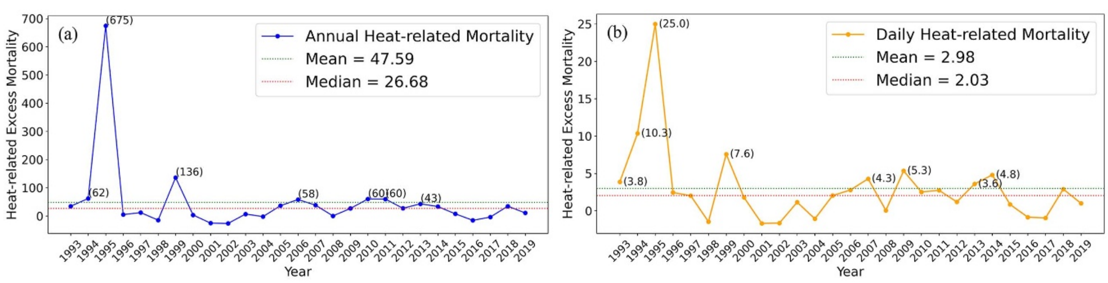

Figure 3. Time series of heat-related excess mortality in Chicago, 1993–2019. (a) Annual heat-related excess mortality; (b) Average daily heat-related excess mortality. Mean and median reference lines are shown, and values exceeding the mean are annotated.

In addition, potential lag effects of heat-related deaths over the study period were assessed. We included the average daily deaths on days 1, 2, and 3 following each heat day and constructed scenarios with different lag periods (lag = 1, 2, and 3 d) for comparison. The results showed no significant differences when the lag effect was considered or not.

This study was approved by the Northwestern University Institutional Review Board (STU00219292).

4. Results

4.1. Heat-related excess mortality

We analyzed the temporal trend of heat days (at least two consecutive days with the daily minimum HI exceeding 70 °F (21.1 °C)) over the 27 year study period (1993–2019) via linear regression. The number of heat days over time shows an upward trend, with an increase of approximately 0.9 heat d per decade, although the trend is not statistically significant ($p = 0.61$). Peak heat day occurrence was in 1995 (27 d) (figure 2).

During the study period, the average heat-related excess mortality in Chicago was ∼48 deaths per year, with ∼3 heat-related deaths per heat day (figure 3). Significant peaks were observed in 1995 and 1999. In 1995, 675 heat-related deaths were estimated, with an average of 25.0 heat-related deaths per heat day; in 1999, there were 136 heat-related deaths and an average of 7.6 heat-related deaths per heat day. Both annual and daily heat-related excess mortality exhibit an overall declining pattern over the study period (figure 3), as illustrated by the linear trend lines shown in supplemental figure S3. To assess whether this long-term trend is disproportionately influenced by the extreme 1995 heatwave in Chicago, we additionally present the time series excluding 1995 in supplemental figures S4 and S5, which show consistent declining trends.

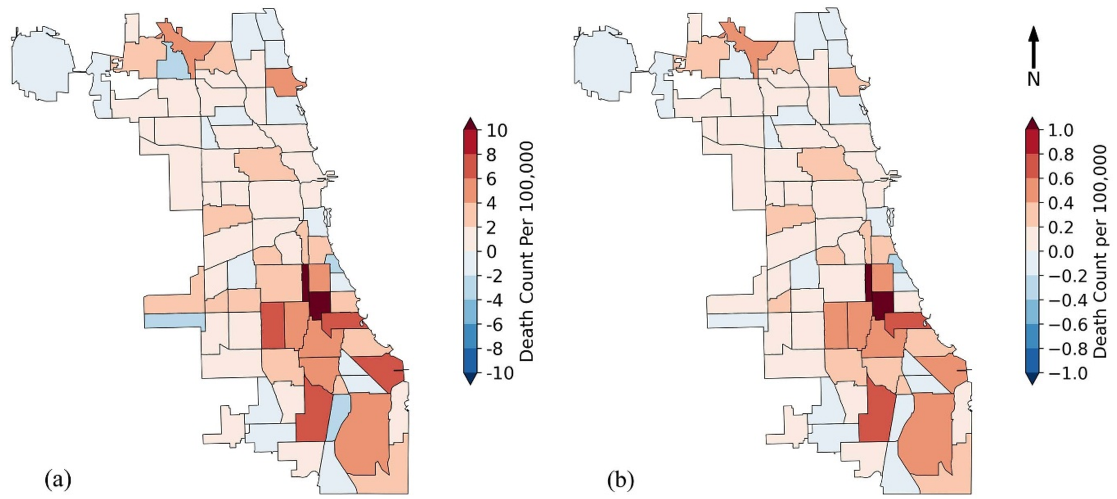

Figure 4. Spatial distribution of heat-related excess mortality rate across communities in Chicago, 1993–2019. Heat days are defined as at least two consecutive days with the daily minimum heat index exceeding 70 °F (21.1 °C). (a) Map of the average annual heat-related excess mortality rate (range: −2.6–13.2 deaths per 100 000 people; mean: 1.8 deaths per 100 000 people). Spatial autocorrelation is statistically significant (Moran's $I = 0.194$, $p = 0.008$). (b) Map of the average daily heat-related excess mortality rate (range: −0.22–1.11 deaths per 100 000 people; mean: 0.16 deaths per 100 000 people). Spatial autocorrelation is statistically significant (Moran's $I = 0.168$, $p = 0.009$).

The spatial distribution of annual and daily average heat-related excess mortality rate across the 77 CAs in Chicago shows that higher values are predominantly concentrated in the southern parts of the city (figure 4). The Moran's I values (0.19 for annual and 0.17 for daily, with both $p < 0.05$) indicate modest but statistically significant positive spatial autocorrelation, suggesting that communities with similar levels of heat-related health risk are often found clustered near one another.

4.2. Variable selection

The heat vulnerability indicators used in the unsupervised HVI, as well as the indicators identified by various selection methods are summarized in table 3. Detailed results for each method are provided in the supplemental materials (supplemental tables S2–S5 and figure S6). Results show that *Age above 65* and *Poverty Rate* are key indicators in all variable selection methods, while *No AC Access* remains an important indicator in four out of the five variable selection methods. *Living Alone* and *Hispanic or Latino (of Any Race)* were not selected by any of the methods, indicating a lack of robust association with heat-related excess mortality in Chicago.

4.3. Principal component analysis

We conducted PCA on two sets of indicators: (1) the full set of candidate heat vulnerability indicators to construct an unsupervised HVI, and (2) separately on each subset of indicators selected by the five supervised methods (SLR, PR, Lasso, RF, XGBoost). The factor loadings, eigenvalues, and proportion of variance explained for all six PCA models are presented in table 3. Differences in the cumulative explained variance across the models primarily reflect variations in the covariance structures of the subsets of indicators retained by each variable selection method.

For the unsupervised HVI, four factors were retained. Factor 1, which explains the maximum variance (35%), is comprised of *Poverty Rate, No AC Access*, and *Black (Non-Hispanic)*. Factor 2 (22.6% of the variance) is characterized by a positive loading for *Less than High School Diploma* and *Hispanic or Latino (of Any Race)*, and a negative loading for *Living Alone*. Factor 3 (12.6% of the variance) shows significant positive loadings for *Age above 65* and *Diabetes Prevalence*. Factor 4 (9.4% of the variance) is dominated by *Not Green Space*.

PCA on the indicators selected by the SLR method identified three factors. Factor 1 (54.3% of the variance) is comprised of *Poverty Rate, No AC Access*, and *Black (Non-Hispanic)*. *Age above 65* and *Age above 65 and Living Alone* form separate factors, explaining 19.6% and 14.0% of the variance, respectively.

PCA on the indicators selected by the PR method also identified three factors. Factor 1 explains 45.4% of the variance and consists of three indicators: *Poverty Rate*, *No AC Access*, and *Black (Non-Hispanic)*. *Less than High School Diploma* and *Age above 65 and Living Alone* together constitute Factor 2, explaining 21.2% of the variance. Notably, these two indicators show opposing loadings on this factor, indicating that their contributions to this factor are in opposite directions, which suggests that communities with a higher proportion of residents lacking a high school diploma tend to have a lower prevalence of adults aged over 65 yr and living alone. Factor 3 (16.2% of the variance) is dominated by *Age above 65*.

PCA on the indicators selected by the Lasso regression method identified four factors. Factor 1, consisting of *Less than High School Diploma* and *Diabetes Prevalence*, explains 32.8% of the variance. The other three factors are dominated by *Age above 65*, *Not Green Space*, and *Poverty Rate*, explaining 23.6%, 18.0%, and 15.6% of the variance, respectively.

PCA on the indicators selected by the RF method extracted three factors. Factor 1 consists of *Poverty Rate* and *No AC Access*, which explains 40.6% of the variance. Factor 2 is similar to the PR result, characterized by opposite loadings between *Less than High School Diploma* and *Age above 65 and Living Alone*. Factor 3 is largely represented by *Age above 65*.

PCA on the indicators selected by the XGBoost method identified three factors. Factor 1, composed of *Poverty Rate, No AC Access*, and *Black (Non-Hispanic)*, explains 41.6% of the variance. Factor 2 loads highly on *Age above 65* and *Diabetes Prevalence* and explains 20.2% of the variance. Factor 3 is dominated by *Not Green Space*, explaining 16.0% of the variance.

Overall, certain indicators consistently loaded strongly on the main factors. The first principal components (Factor 1), which explain maximum variance, always include *Poverty Rate* and *No AC Access*, and sometimes also include *Black (Non-Hispanic)*. The only exception occurs in the PCA results using indicators selected by Lasso regression, where neither *Poverty Rate* nor *No AC Access* appear in Factor 1, but *Poverty Rate* loads separately on Factor 4, explaining 16% of the variance. This factor generally represents a socioeconomic vulnerability dimension associated with adaptive capacity and shows high stability across different methods, highlighting its prominent role in Chicago's heat vulnerability. *Age above 65* is also a robust indicator, loading significantly on at least one factor in all methods. The elderly population is regarded as a key indicator of sensitivity in heat vulnerability, and its consistent inclusion across all methods further underscores the importance of sensitivity-related characteristics. In addition, *Not Green Space* often forms a separate single factor, indicating that as an environmental variable, it exhibits a relatively independent pattern of variation and reflects the distinct contribution of environmental exposure to heat vulnerability. Although the factor structures obtained using different variable selection methods vary, the adaptive capacity, sensitivity, and exposure dimensions of heat vulnerability are consistently represented in the results.

4.4. Construction and evaluation of heat vulnerability indices

HVIs constructed from the different variable selection methods have distinct spatial patterns but predominantly identify the southern and western CAs as those that are most vulnerable to heat (figure 5). The unsupervised HVI (figure 5(a)) shows that areas with higher vulnerability are mainly concentrated in the southern and western CAs. The SLR model produces HVI scores with higher values mainly concentrated in the southeastern CAs (figure 5(b)). The PR-informed HVI (figure 5(c)) indicates higher heat vulnerability in the southern and western CAs. The Lasso model generates HVI scores (figure 5(d)) showing a general trend of higher vulnerability in the west and south. The RF and XGBoost models likewise show the highest HVI scores concentrated in the southern-and western regions of the city (figure 5(e & f)).

Since each PCA was performed on a different subset of indicators, the variance structure explained and the number of retained factors differ. As a result, the absolute HVI scores are on different numerical scales and cannot be directly compared across methods. To enable cross-method comparison, we transformed the HVIs to a unified ordinal scale by categorizing scores into Levels 1–4 using equal width binning. Figure 6 presents the spatial distribution of resulting heat vulnerability levels. Both the unsupervised HVI (figure 6(a)) and the XGBoost-informed HVI (figure 6(f)) identify scattered high vulnerability CAs (Level 4). The other methods identify fewer and more spatially clustered high heat vulnerability areas, with overlap among models in several southern CAs (figures 6(b)–(e)). For the medium-high heat vulnerability areas (Level 3), most maps show a band-like distribution extending from the northwest to the southeast, with the exception of the SLR-informed method, which exhibits higher heat vulnerability in eastern CAs (figure 6(b)). For medium-low heat vulnerability areas (Level 2), the distribution generally surrounds high and medium-high vulnerability CAs, without obvious spatial clustering. Low vulnerability areas (Level 1) are relatively uniformly distributed, primarily along the northern and southern margins of the city, as well as in several CAs near Lake Michigan in the northeast. In summary, despite variations in the precise location and extent of the highest-risk clusters, the HVI maps converge on a common pattern in which Chicago's northeastern CAs generally exhibit lower heat vulnerability, while higher vulnerability areas are mostly concentrated in the southern and western parts of the city.

Table 3. Variable selection and principal component analysis results. Indicators used to construct the unsupervised heat vulnerability index (HVI) and those retained by each variable selection method (simple linear regression (SLR), polynomial regression (PR), Lasso regression, random forest (RF), and XGBoost) for the supervised HVIs are presented under the corresponding method. Blank cells indicate indicators not selected by a given method. Factor loadings for heat vulnerability indicators are presented along with eigenvalues and the proportion of variance explained by each factor. Absolute values > 0.4 are considered the most significant loadings on a respective factor and are marked in bold.

	Unsupervised HVI				Simple Linear Regression			Polynomial Regression			Lasso Regression				Random Forest			XGBoost		
Indicator	Factor 1	Factor 2	Factor 3	Factor 4	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3	Factor 4	Factor 1	Factor 2	Factor 3	Factor 1	Factor 2	Factor 3
Not Green Space	−0.05	−0.01	0.02	**0.91**							0.00	0.01	**1.00**	0.00				0.00	−0.01	**0.94**
Age above 65	−0.07	−0.15	**0.74**	−0.11	0.01	**0.97**	−0.01	0.00	0.04	**0.96**	0.00	**0.91**	0.01	0.06	0.02	−0.01	**0.99**	0.03	**0.69**	−0.18
Age above 65 and Living Alone	0.24	−0.35	−0.02	0.21	0.08	0.00	**0.91**	0.29	**−0.58**	−0.10					0.39	**−0.63**	−0.02			
Living Alone	0.02	**−0.51**	−0.01	0.20																
Black (Non-Hispanic)	**0.47**	−0.15	0.15	−0.09	**0.51**	0.20	0.10	**0.46**	−0.22	0.23								**0.57**	0.13	−0.06
Hispanic or Latino (of Any Race)	−0.09	**0.51**	−0.05	0.16																
Diabetes Prevalence	0.07	0.18	**0.63**	0.12							**0.70**	0.30	−0.02	−0.17				−0.01	**0.70**	0.18
Less than High School Diploma	0.27	**0.53**	0.08	0.10				0.30	**0.78**	−0.06	**0.71**	−0.29	0.02	0.17	0.31	**0.77**	−0.02			
Poverty Rate	**0.63**	0.04	−0.14	−0.09	**0.54**	−0.14	0.20	**0.59**	−0.01	−0.13	0.00	0.05	0.00	**0.97**	**0.65**	−0.03	−0.08	**0.63**	−0.13	−0.09
No AC Access	**0.49**	0.05	0.00	0.12	**0.67**	−0.05	−0.35	**0.51**	0.09	0.04					**0.57**	0.05	0.09	**0.52**	0.00	0.19
Eigenvalue	3.55	2.29	1.27	0.95	2.75	0.99	0.71	2.76	1.29	0.98	1.66	1.19	0.91	0.79	2.06	1.26	0.97	2.53	1.23	0.97
% Variance Explained	35.0	22.6	12.6	9.4	54.3	19.6	14.0	45.4	21.2	16.2	32.8	23.6	18.0	15.6	40.6	24.9	19.2	41.6	20.2	16.0

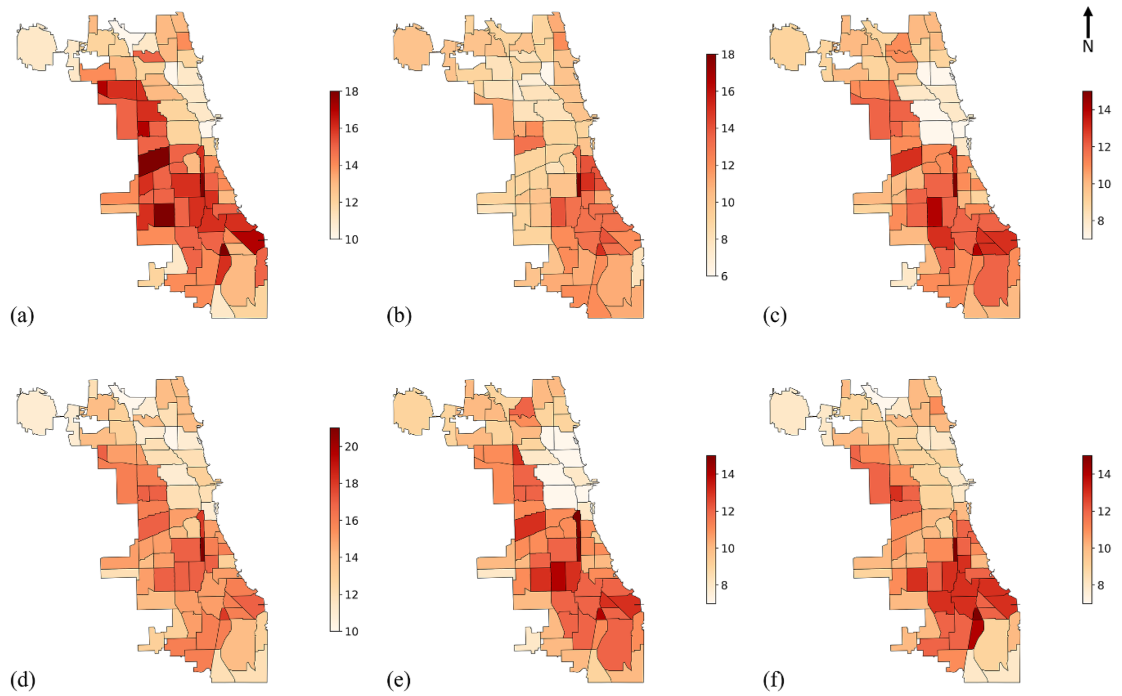

Figure 5. Heat vulnerability index (HVI) maps. HVI scores were generated using indicator subsets selected by different methods: (a) Unsupervised HVI; (b) Simple linear regression; (c) Polynomial regression; (d) Lasso regression; (e) Random forest; (f) XGBoost. Higher HVI scores indicate higher heat vulnerability.

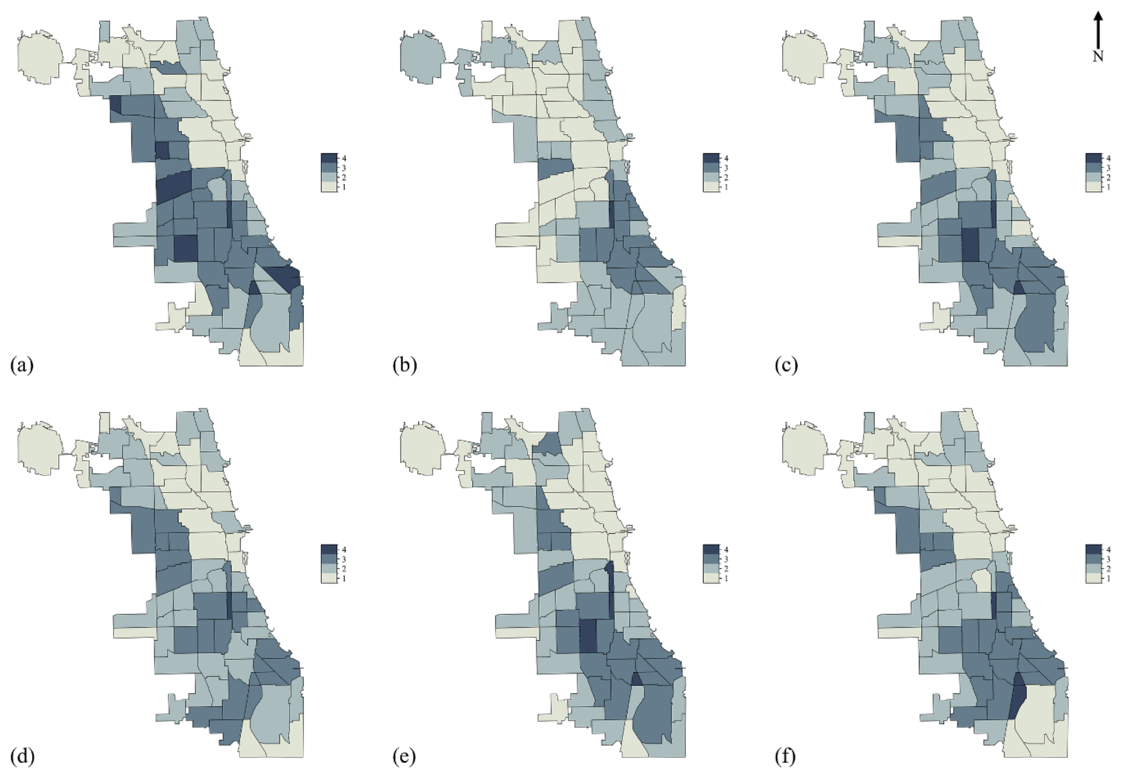

Figure 6. Heat vulnerability index (HVI) category maps. HVI scores were generated using indicator subsets selected by different methods: (a) Unsupervised HVI; (b) Simple linear regression; (c) Polynomial regression; (d) Lasso regression; (e) Random forest; (f) XGBoost. The HVI scores were classified into four categories using equal width binning: 1 = Low, 2 = Medium-low, 3 = Medium-high, 4 = High. Higher levels indicate higher heat vulnerability.

Table 4. Performance metrics of different heat vulnerability index (HVI) models for continuous and categorical tasks. Spearman's ρ: Spearman's rank correlation coefficient; higher values indicate stronger monotonic agreement between predicted heat vulnerability and observed heat-related excess mortality. Mean absolute error (MAE) and mean squared error (MSE): Lower values indicate better model performance. Accuracy: Proportion of correct classifications in categorical tasks; higher values indicate better performance. *F*1 score: Harmonic mean of precision and recall in categorical tasks; higher values indicate better balance between precision and recall. Best-performing values for each metric are shown in bold.

	Continuous Metrics			Categorical Metrics	
HVI Models	Spearman's ρ	MAE	MSE	Accuracy	*F*1 score
Unsupervised HVI	0.29	13.87	196.13	0.32	0.35
Simple Linear Regression	0.28	10.41	112.92	0.42	0.42
Polynomial Regression	0.32	**10.39**	**110.69**	0.47	0.48
Lasso Regression	0.30	14.23	206.79	0.45	0.48
Random Forest	**0.37**	10.41	111.61	**0.49**	**0.51**
XGBoost	0.28	10.45	112.08	0.45	0.46

To evaluate the performance of each HVI, we assessed the correlation and categorical classification metrics between HVI scores and estimated heat-related excess mortality for each of the 77 CAs. The results of both continuous evaluation metrics (Spearman correlation, MAE, and MSE) and categorical classification metrics (accuracy and *F*1 score) are summarized in table 4. For continuous evaluation, the RF-informed HVI exhibited the highest Spearman correlation coefficient with heat-related excess mortality ($\rho = 0.37$), exceeding that of the unsupervised HVI ($\rho = 0.29$) and the SLR-informed HVI ($\rho = 0.28$). This level of correlation suggests a moderate monotonic association. In addition, compared with the other HVIs, the RF-informed HVI demonstrated relatively lower MAE and MSE values, indicating its superior overall performance under continuous evaluation metrics. In categorical evaluation, the RF-informed HVI achieves an accuracy of 0.49 and *F*1 score of 0.51, both of which are the best among the evaluated HVIs. These values are 0.17 and 0.16 higher than those of the unsupervised HVI (Accuracy = 0.32, *F*1 score = 0.35), corresponding to relative increases of approximately 53% and 46%.

The PR-informed HVI achieves the lowest MAE and MSE among all models, while its Spearman correlation and classification accuracy are slightly lower than those of the RF-informed HVI. In continuous prediction, the HVIs informed by SLR and XGBoost have the poorest Spearman correlation values among methods despite having lower MAE and MSE, whereas the Lasso-informed HVI produces the highest MAE and MSE. In categorical prediction, the SLR-informed HVI shows an improvement over the unsupervised HVI but remains inferior to both RF- and PR-informed HVIs (table 4).

4.5. Sensitivity analyses

To examine the robustness of several of our assumptions, we conducted sensitivity analyses using different analytical settings, including: (1) alternative heat day definitions; (2) age-standardized heat-related mortality data; and (3) lag effects of heat-related mortality, with corresponding maps shown in supplemental figures S7 and S8.

These analyses suggest that the key heat vulnerability indicators are robust across analytical settings. That is, regardless of extreme heat definition or age standardization, *Poverty Rate* and *No AC Access* are consistently identified as important indicators by most variable selection methods. *Age above 65* also remained highly selected, particularly under nighttime heat definitions, although its relative ranking showed modest variation across analytical settings. *Black (Non-Hispanic)*, *Less than High School Diploma*, and *Age above 65 and Living Alone* show higher selection frequency compared with the other indicators (figure 7).

5. Discussion

This study systematically compares the performance of various variable selection methods in the construction of PCA-based HVIs, using a set of established heat vulnerability indicators derived from the published literature. The results indicate that, under our study design and data conditions, *No AC Access, Poverty Rate,* and *Age above 65* at the community level consistently emerged as important determinants of heat vulnerability in Chicago. RF-based variable selection outperforms other methods when the results are evaluated against heat-related excess mortality data, suggesting that supervised machine learning-based variable selection holds great promise for improving heat vulnerability assessments.

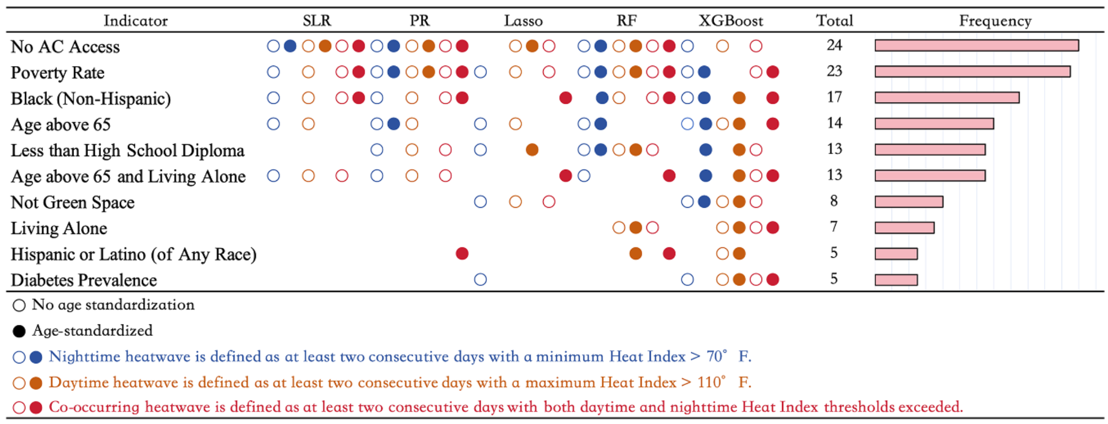

Figure 7. Variable selection results across different analytical configurations. Two configurations were considered: (1) weather indicator defined as daily maximum heat index > 110 °F (43.3 °C), daily minimum heat index > 70 °F (21.1 °C), or co-occurring exceedance of daytime and nighttime heat index thresholds; and (2) age-standardized vs. non-standardized heat-related mortality. The bar plot shows the selection count of each indicator, with indicators ordered by selection frequency. AC = Air conditioning; SLR = Simple linear regression; PR = Polynomial regression; RF = Random forest.

5.1. Consistency and variability of determinants for heat-related health risks

The identification of *Poverty Rate* and *No AC Access* in our analysis aligns with previous retrospective analyses of Chicago heatwaves. During the 1995 Chicago heatwave, deaths were more prevalent in the low-income areas of the city, and neighborhood affluence was negatively associated with heatwave mortality (Klinenberg 1999, Browning *et al* 2006). In both the 1995 and 1999 heatwaves, access to AC was consistently associated with lower death risk and was identified as a strong protective factor (Semenza *et al* 1996, Naughton *et al* 2002). Our results also align with broader global patterns. Existing empirical studies in multiple countries and cities have shown that, under heat exposure, economically disadvantaged populations often exhibit higher mortality rates or disease burdens (Greenberg *et al* 1983, Curriero *et al* 2002, Chan *et al* 2012, Fletcher *et al* 2012, Madrigano *et al* 2013, Zanobetti *et al* 2013, Toloo *et al* 2014, Lim *et al* 2015, de Visser *et al* 2022), while an increase in AC prevalence has been linked to lower heat-related mortality and hospitalization rates (Greenberg *et al* 1983, Curriero *et al* 2002, Davis *et al* 2003, O'Neill *et al* 2005, Bouchama *et al* 2007, Anderson and Bell 2009, Ostro *et al* 2010).

In addition to the vulnerability indicators discussed above, older age, race, and educational attainment were also frequently identified as important indicators in this study. During the 1995 Chicago heatwave, Black residents and older adults faced a disproportionately higher risk of death (Whitman *et al* 1997, Klinenberg 1999, 2022, Semenza *et al* 1999, Naughton *et al* 2002, Kaiser *et al* 2007). These early findings are further corroborated by more recent community-level research on heat-related cardiovascular morbidity and mortality in Chicago, which identifies lower income, unemployment, lower educational attainment, and a high proportion of Black residents as dominant contributors to structural vulnerability (Graffy *et al* 2025). These factors have also been recognized by extensive epidemiological studies as being significantly associated with higher heat-related health risks (e.g. Curriero *et al* 2002, O'Neill *et al* 2003, Anderson and Bell 2009, Fletcher *et al* 2012, Zanobetti *et al* 2013, Kephart *et al* 2022).

Some of our findings diverge from those reported in previous studies. For instance, *Living Alone* did not show consistent importance across various variable selection methods in our research. However, case-control studies reported that individuals living alone faced a higher risk of death during the 1995 and 1999 Chicago heatwaves (Semenza *et al* 1996, Naughton *et al* 2002). This difference may be related to the research scale. Klinenberg (1999) pointed out that this individual-level methodological approach leaves the differences between CAs invisible. Although living alone is an important individual risk factor, it may not serve as a robust predictor of heat-related mortality at the community level. Browning *et al* (2006) found in a neighborhood-level quantitative analysis that while the proportion of residents living alone was significantly and positively associated with the baseline mortality, its explanatory power for the variation in mortality during the 1995 Chicago heatwave was not significant. In addition, we identified the proportion of Black residents as an important heat vulnerability indicator, whereas some prior studies did not find a statistically significant association between race and certain heat-related health outcomes. Green *et al* (2010) reported no statistically significant effect modification by race on temperature-associated hospital admissions in nine counties in California; similarly, Pillai *et al* (2014) found no significant differences between White and Black individuals in terms of the proportion of heat-related emergency department visits resulting in hospital admission (versus routine discharge) in Georgia. Shannon *et al* (2025) found that extreme heat was associated with increased mortality among California veterans with cardiometabolic diseases, but reported no statistically significant effect modification by race. These authors attributed the absence of observed race effects primarily to differences in outcome definitions and institutional contexts. Specifically, hospital admissions may not fully capture heat-related morbidity across racial groups because of possible differences in health care utilization (Green *et al* 2010), whereas the integrated Veterans Affairs care delivery paradigm may reduce racial disparities through targeted interventions addressing social factors related to health (Shannon *et al* 2025). These examples collectively suggest that race does not always exhibit a stable association with heat-related health outcomes across various settings, highlighting the strong context dependence of its importance as a heat vulnerability indicator.

5.2. Heterogeneity and context-dependence of heat vulnerability

As regional contexts differ in population structure, health care systems, and health insurance coverage, heat vulnerability exhibits significant spatial heterogeneity. People in different regions respond differently to heat exposure (Gasparrini *et al* 2015). Consequently, the key predictors for heat-related health outcomes can vary across regions (Anderson and Bell 2009, Arsad *et al* 2022). For example, a study by Mallen *et al* (2019) in Dallas, Texas, found that older populations, people without a high school diploma, and lack of access to AC were important indicators of heat-related mortality. Whereas in Detroit, Michigan, Conlon *et al* (2020) found that older age, socioeconomic status, race, and ethnicity indicators were not robust stand-alone determinants of heat vulnerability. In Phoenix, Arizona, Chuang and Gober (2015) found that lower socioeconomic status, the proportion of adults aged over 65 years and living alone, the proportion of adults living alone, and the diabetes hospitalization rate could be used to predict heat vulnerability at the census tract level, while AC prevalence was not a significant predictor of heat-related health outcomes. Studies focused on non-U.S. locales show similar nuance. Li *et al* (2025), in their study of heat vulnerability in Australian capital cities, suggested that personal health status and sociodemographic characteristics (including individual disease status, age, and education level) play a dominant role in determining heat vulnerability, overshadowing the influence of environmental and infrastructure factors. In contrast, Jeong *et al* (2024), in a city-level study in South Korea, found that climate conditions and income among the elderly population showed statistically significant associations with heat-related mortality. These case studies from diverse locales indicate that the explanatory power of specific heat vulnerability indicators can be inconsistent across different geographic contexts, with each city exhibiting a distinct combination of important indicators.

These local nuances highlight that there is no single, stable set of heat vulnerability indicators that can be universally applied across different regional settings. Differences in climatic conditions, population structure, socioeconomic composition, and health care systems jointly shape distinct local heat-related health risk mechanisms across unique cities and regions. Given this context, our study emphasizes the importance of adopting supervised HVI frameworks where variable selection is informed by local heat-related health outcomes. Compared with unsupervised approaches, this data-driven step allows for the statistical identification of core combinations of indicators that are most explanatory of local risks, leading to more accurate heat vulnerability assessments that can better inform the allocation of resources and targeted intervention strategies. Practically, this supports avoiding one-size-fits-all HVIs and instead selecting and validating indicators against local outcomes and community knowledge prior to using maps to focus interventions (Harris *et al* 2018, Potash *et al* 2020).

5.3. Methodological differences in supervised heat vulnerability index construction

In this context, the choice of variable selection method can directly influence HVI outcomes, as different methods are grounded in distinct statistical assumptions and algorithmic mechanisms. Traditional SLR focuses on the linear relationship between the outcome variable (in our case, heat-related excess mortality) and each heat vulnerability indicator, and thus, less effectively captures potential nonlinear effects or complex interactions among indicators. In contrast, PR can fit a certain degree of nonlinear relationships by introducing higher order terms. In our variable selection results, PR identified the *Less than High School Diploma* indicator along with a set of indicators largely consistent with those identified by SLR. This result suggests a potential nonlinear association between educational attainment and heat-related excess mortality, where heat vulnerability may increase disproportionately once the proportion of residents with low educational attainment exceeds a certain level. However, PR is still based on a linear model framework with fixed parameter forms, which limits its ability to fully capture more complex nonlinear relationships and interactions.

As a machine learning approach, Lasso regression is also sensitive to linear relationships and tends to retain a representative variable in a highly correlated group while setting the coefficients of others to zero. This mechanism could lead to the omission of variables with practical significance. In this study, the AC access indicator was robustly identified by other methods but not selected by Lasso regression. This is likely attributable to its moderate correlation with poverty rates ($r = 0.59$, see supplemental figure S2). The indicators retained by Lasso regression exhibited relatively weak overall linear correlations, resulting in the variance being distributed across multiple principal components in the subsequent PCA rather than being concentrated in the first principal component. Both RF and XGBoost are capable of capturing nonlinear effects. Their selection mechanisms may retain indicators that jointly represent the same latent dimension, allowing a greater proportion of the total variance to be explained by the first few principal components. Although XGBoost is known for strong capabilities in capturing feature interactions and nonlinear signals, it underperformed across evaluation metrics in our study. This may not reflect a limitation of the XGBoost algorithm itself, but rather its heightened sensitivity to noise under the small sample size ($n = 77$ CAs). The RF-informed HVI most accurately depicted the distribution of heat-related mortality, demonstrating the potential of this approach for developing precise heat vulnerability assessment tools under the conditions of this study. Overall, the performance of any variable selection method in this study should not be generalized across all contexts. Rather than overinterpreting results from a single approach, greater emphasis should be placed on indicators that demonstrate stable selection across multiple methods.

5.4. Implications for localized heat vulnerability governance

For implementation, these HVIs should be viewed as decision-support tools rather than sole eligibility screens. They are best used alongside local practitioner insight and lived experience and should be revisited as housing conditions, cooling access, demographics, and climate hazards change. Where feasible, localities should also pair HVI-based prioritization with monitoring of service uptake and health outcomes to reduce the risk of systematic under-identification of vulnerable groups. Our findings suggest that Chicago's heat risk management and resource allocation efforts could prioritize communities with higher levels of poverty. For these communities, improving access to AC, expanding cooling assistance programs, and increasing the number of cooling centers may be helpful to reduce heat exposure.

While it is notable that our study identifies RF as the best performing supervised variable selection method, the resultant Spearman's correlation coefficient (0.37) and classification accuracy of less than 50% leave ample room for improvement. One reason for the muted performance may be that we limited our candidate heat vulnerability indicators to those used by Reid *et al* (2009). The choice was made intentionally to assess the different HVI construction approaches rather than to create a high-fidelity characterization of Chicago's heat vulnerability. Since this candidate indicator set may not include certain key determinants of heat vulnerability specific to the Chicago context, the resulting HVIs likely only partially reflect the complexity of heat-related health risks in the city.

For higher-fidelity, location-specific applications, researchers should consider incorporating a comprehensive set of candidate indicators. The systematic review by Cheng *et al* (2021) on HVIs emphasizes the importance of including as many relevant indicators as possible. Alternatively (or complementary), a community-informed approach to indicator identification could be used to select a broader set of locally-informed and contextually-relevant indicators; a methodological choice designed to leverage the knowledge of lived experience and to build trust amongst the community of users (Defusing Disasters Working Group 2026).

5.5. Limitations

Existing research has shown that HVIs can be sensitive to the scale, measurement, and institutional and social context in which indicators are applied (Hinkel 2011, Tate 2012, Chuang and Gober 2015, Conlon *et al* 2020). For instance, living alone is often interpreted as reflecting vulnerability arising from insufficient social support in heat vulnerability research, and conceptually is more regarded as a proxy variable for social isolation (Klinenberg 1999). However, the strength of an individual's social support network is not entirely determined by whether they are living alone. At the contemporary community scale in Chicago, the contributory effect of living alone on heat vulnerability may be overshadowed by more dominant structural factors such as poverty rate and lack of AC. Therefore, the statistical importance of living alone may vary across research contexts and analysis scales.

Additionally, spatiotemporal limitations of the data may also constrain the ability of the HVIs to characterize heat vulnerability. Limited temporal coverage of demographic, socioeconomic, and health data often necessitates the assumption that vulnerability indicators change uniformly over time, i.e., that the relative differences between communities remain constant throughout the study period. Though this assumption is common in heat vulnerability research, it may lead to biased estimation, especially over longer study periods or when using temporally dynamic indicators. For the land cover indicator *No Green Space*, averaging the data over decades may not fully capture the instantaneous exposures of the population when areas change rapidly with tree or vegetation loss due to urban development or other causes. The spatially smoothed Daymet dataset was derived through grid interpolation and may not accurately represent temperature extremes in urban microclimates, potentially leading to an underestimation of heat exposure in areas with intense heat accumulation in cities like Chicago.

Beyond data constraints, the thresholds and metrics used to characterize heat exposure may also influence the assessment. This study adopted 70 °F (approximately 21.1 °C) as the warm-night threshold. The health impact estimates associated with the heat events identified by this threshold are generally consistent with those reported in existing studies on historical extreme heat events in Chicago (Semenza *et al* 1996, Whitman *et al* 1997, Palecki *et al* 2001, Kaiser *et al* 2007). However, this threshold may not fully reflect the local acclimatization patterns and climatic context of the population in Chicago and the Midwestern U.S. Future research should further explore locally calibrated temperature thresholds, thereby establishing more context-specific definitions and assessment criteria for heat exposure. Furthermore, although the original heat d threshold was defined based solely on air temperature, we employed HI as the exposure metric to better reflect the influence of humidity on human thermal perception. This choice of heat stress metric may also influence health impact assessments, as different metrics capture unique dimensions of heat stress, leading to varying estimates of heat-related health impacts (Schwingshackl *et al* 2021).

6. Conclusion

This study compared the performance of unsupervised and supervised variable selection methods for the construction of HVIs. Under the settings and data conditions of this study, a RF-informed HVI exhibited the most robust performance, highlighting the potential of machine learning approaches to capture the nonlinear relationships between heat vulnerability indicators and heat-related health outcomes.

The findings further suggest that traditional unsupervised methods for constructing HVIs may have inherent limitations. These methods lack rigorous data-driven variable selection validated against health outcomes, often rely on predefined indicator sets, and implicitly assume that the determinants of heat vulnerability are consistent across different regions. Heat vulnerability is context-dependent and shaped by specific social, environmental and climatic conditions. Therefore, neither indicator sets nor the performance of a given methodological approach can be assumed to be directly transferable across distinct geographic areas.

Based on this understanding, the main contribution of this study lies in presenting a health outcome-informed framework for HVI construction. Within this framework, Chicago serves as a case study to demonstrate how modeling and interpretation can be conducted using local high-resolution data. In heat vulnerability assessments, variable selection strategies should be systematically evaluated in accordance with specific research objectives, regional context, and data characteristics. Locally tailored, outcome-informed validation is therefore essential before HVI maps are used to inform equitable resource allocation for climate adaptation and heat risk mitigation. This also highlights the need for other localities to collect and make available high-resolution data in order to carry out effective heat risk assessments within their own unique contexts.

Acknowledgment

This work was sponsored by the School of Integrated Climate and Earth System Sciences at University of Hamburg; and the Office of Global Initiatives at the McCormick School of Engineering, the Paula M. Trienens Institute for Sustainability and Energy, and the Roberta Buffett Institute for Global Affairs at Northwestern University. We thank the Illinois and Chicago Departments of Public Health for providing the mortality records. JS, BP, LB and MvS acknowledge funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—EXC 2037 'CLICCS—Climate, Climatic Change, and Society'—Project Number 390683824.

Data availability statement

The data that support the findings of this study are available upon reasonable request to the authors. The meteorological and socioeconomic data used in this study are publicly available. However, the mortality data are owned by a third party and contain sensitive health information; legal and privacy restrictions prevent their public distribution. These data may be available from the original data providers upon reasonable request.

Supplementary Information available at: https://doi.org/10.1088/2752-5309/ae9633/data1.

Author contributions

Shuyue Qu 0009-0009-2752-9741
Conceptualization (equal), Data curation (equal), Formal analysis (equal), Investigation (equal), Methodology (equal), Software (equal), Validation (equal), Visualization (equal), Writing – original draft (equal), Writing – review & editing (equal)

Jana Sillmann 0000-0002-0219-5345
Conceptualization (equal), Funding acquisition (equal), Methodology (equal), Project administration (equal), Resources (equal), Supervision (equal), Writing – review & editing (equal)

Benjamin W Barrett 0000-0002-3292-7017
Conceptualization (equal), Data curation (equal), Methodology (equal), Software (equal), Writing – review & editing (equal)

Peter M Graffy 0000-0002-2156-7175
Conceptualization (equal), Data curation (equal), Methodology (equal), Software (equal), Writing – review & editing (equal)

Benjamin Poschlod 0000-0003-0247-2514
Methodology (equal), Software (equal), Writing – review & editing (equal)

Lukas Brunner 0000-0001-5760-4524
Methodology (equal), Writing – review & editing (equal)

Raed Mansour 0000-0002-8484-0730
Conceptualization (equal), Resources (equal), Writing – review & editing (equal)

Malte von Szombathely 0000-0002-3300-7482
Writing – review & editing (equal)

Finley Hay-Chapman 0000-0002-4161-8494
Writing – review & editing (equal)

Teresa H Horton 0000-0002-3734-8065
Conceptualization (equal), Writing – review & editing (equal)

Jennifer Chan 0000-0001-7184-6335
Conceptualization (equal), Writing – review & editing (equal)

Sheetal Khedkar Rao 0000-0003-2763-7741
Conceptualization (equal), Writing – review & editing (equal)

Kyra Woods
Conceptualization (equal), Writing – review & editing (equal)

Abel N Kho 0000-0003-1993-5634
Supervision (equal), Writing – review & editing (equal)

Daniel E Horton 0000-0002-2065-4517
Conceptualization (equal), Funding acquisition (equal), Methodology (equal), Project administration (equal), Resources (equal), Supervision (equal), Writing – review & editing (equal)

References

Anderson B G and Bell M L 2009 Weather-related mortality: how heat, cold, and heat waves affect mortality in the United States *Epidemiology* **20** 205–13

Arsad F S, Hod R, Ahmad N, Ismail R, Mohamed N, Baharom M, Osman Y, Radi M F M and Tangang F 2022 The impact of heatwaves on mortality and morbidity and the associated vulnerability factors: a systematic review *Int. J. Environ. Res. Public Health* **19** 16356

Azhar G S, Mavalankar D, Nori-Sarma A, Rajiva A, Dutta P, Jaiswal A and Hess J J 2014 Heat-related mortality in India: excess all-cause mortality associated with the 2010 Ahmedabad heat wave *PLoS One* **9** e91831

Barnett A G and Åström C 2012 Commentary: what measure of temperature is the best predictor of mortality? *Environ. Res.* **118** 149–51

Bell M L, Gasparrini A and Benjamin G C 2024 Climate change, extreme heat, and health *New Engl. J. Med.* **390** 1793–801

Benmarhnia T, Deguen S, Kaufman J S and Smargiassi A 2015 Vulnerability to heat-related mortality: a systematic review, meta-analysis, and meta-regression analysis *Epidemiology* **26** 781–93

Bouchama A, Matthies F, Shoukri M, Dehbi M, Mohamed G and Menne B 2007 Prognostic factors in heat wave–related deaths: a meta-analysis *Arch. Intern. Med.* **167** 2170–6

Breiman L 2001 Random forests *Mach. Learn.* **45** 5–32

Browning C R, Wallace D, Feinberg S L and Cagney K A 2006 Neighborhood social processes, physical conditions, and disaster-related mortality: the case of the 1995 Chicago heat wave *Am. Sociol. Rev.* **71** 661–78

Cai J, Luo J, Wang S and Yang S 2018 Feature selection in machine learning: a new perspective *Neurocomputing* **300** 70–79

Chan E Y Y, Goggins W B, Kim J J and Griffiths S M 2012 A study of intracity variation of temperature-related mortality and socioeconomic status among the Chinese population in Hong Kong *J. Epidemiol. Community Health* **66** 322–7

Chen T and Guestrin C 2016 Xgboost: a scalable tree boosting system *Proc. 22nd Acm Sigkdd Int. Conf. on Knowledge Discovery and Data Mining* pp 785–94

Cheng W, Li D, Liu Z and Brown R D 2021 Approaches for identifying heat-vulnerable populations and locations: a systematic review *Sci. Total Environ.* **799** 149417

Chuang W C and Gober P 2015 Predicting hospitalization for heat-related illness at the census-tract level: accuracy of a generic heat vulnerability index in phoenix, Arizona (USA) *Environ. Health Perspect.* **123** 606–12

Conlon K C, Mallen E, Gronlund C J, Berrocal V J, Larsen L and O'Neill M S 2020 Mapping human vulnerability to extreme heat: a critical assessment of heat vulnerability indices created using principal components analysis *Environ. Health Perspect.* **128** 097001

Curriero F C, Heiner K S, Samet J M, Zeger S L, Strug L and Patz J A 2002 Temperature and mortality in 11 cities of the eastern united states *Am. J. Epidemiol.* **155** 80–87

Davis R E, Knappenberger P C, Michaels P J and Novicoff W M 2003 Changing heat-related mortality in the united states *Environ. Health Perspect.* **111** 1712

de Visser M, Kunst A E and Fleischmann M 2022 Geographic and socioeconomic differences in heat-related mortality among the Dutch population: a time series analysis *BMJ Open* **12** e058185

Defusing Disasters Working Group 2026 *Building a Heat-Resilient Chicago: A blueprint for community and City collaboration to stay safe now and prepare for a hotter future*, Northwestern University Roberta Buffett Institute for Global Affairs

Ebi K L, Capon A, Berry P, Broderick C, de Dear R, Havenith G and Jay O 2021 Hot weather and heat extremes: health risks *Lancet.* **398** 698–708

Emmert-Streib F and Dehmer M 2019 High-dimensional LASSO-based computational regression models: regularization, shrinkage, and selection *Mach. Learn. Knowl. Extr.* **1** 359–83

Fletcher B A, Lin S, Fitzgerald E F and Hwang S A 2012 Association of summer temperatures with hospital admissions for renal diseases in New York State: a case-crossover study *Am. J. Epidemiol.* **175** 907–16

Fouillet A, Rey G, Laurent F, Pavillon G, Bellec S, Guihenneuc-Jouyaux C and Hémon D 2006 Excess mortality related to the august 2003 heat wave in France *Int. Arch. Occup. Environ. Health* **80** 16–24

Gasparrini A, Guo Y, Hashizume M, Lavigne E, Zanobetti A, Schwartz J and Armstrong B 2015 Mortality risk attributable to high and low ambient temperature: a multicountry observational study *Lancet.* **386** 369–75

Genuer R, Poggi J M and Tuleau-Malot C 2010 Variable selection using random forests *Pattern Recognit. Lett.* **31** 2225–36

Graffy P M, Barrett B W, Horton D E, Allen N B and Kho A N 2026 Heat and hearts: an exposure-anchored computational phenotyping framework for assessing cardiovascular vulnerability during extreme heat *J. Biomed. Inform.* **22** 105029

Graffy P M, Barrett B W, Horton D E, Kho A N and Allen N B 2025 Critical temperature thresholds for identifying vulnerability to heat-related excess cardiovascular morbidity and mortality *J. Am. Heart Assoc.* ***15*** e046117

Graffy P M, Sunderraj A, Visa M A, Miller C H, Barrett B W, Rao S and Horton D E 2024 Methodological approaches for measuring the association between heat exposure and health outcomes: a comprehensive global scoping review *GeoHealth* **8** e2024GH001071

Green R S, Basu R, Malig B, Broadwin R, Kim J J and Ostro B 2010 The effect of temperature on hospital admissions in nine California counties *Int J Public Health* **55** 113–21

Greenberg J H, Bromberg J I L L, Reed C M, Gustafson T L and Beauchamp R A 1983 The epidemiology of heat-related deaths, Texas–1950, 1970–79, and 1980 *Am. J. Public Health* **73** 805–7

Grömping U 2009 Variable importance assessment in regression: linear regression versus random forest *Am. Stat.* **63** 308–19

Guo W and Zhou Z Z 2022 A comparative study of combining tree-based feature selection methods and classifiers in personal loan default prediction *J. Forecast.* **41** 1248–313

Harris J K, Hinyard L, Beatty K, Hawkins J B, Nsoesie E O, Mansour R and Brownstein J S 2018 Evaluating the implementation of a twitter-based foodborne illness reporting tool in the city of St. Louis department of health *Int. J. Environ. Res. Public Health* **15** 833

Hastie T, Tibshirani R and Friedman J 2009 The elements of statistical learning

Hinkel J 2011 "Indicators of vulnerability and adaptive capacity": towards a clarification of the science–policy interface *Glob. Environ. Change.* **21** 198–208

IPCC 2023 Climate change 2023: synthesis report ed H Lee and J Romero (Core Writing Team) *Contribution of Working Groups I, II and III to the 6th Assessment Report of the Intergovernmental Panel on Climate Change* (IPCC) pp 35–115

Jeong S, Lim Y, Kang Y and Yi C 2024 Elucidating uncertainty in heat vulnerability mapping: perspectives on impact variables and modeling approaches *Int. J. Environ. Res. Public Health* **21** 815

Jiang Z, Che J, He M and Yuan F 2023 A CGRU multi-step wind speed forecasting model based on multi-label specific XGBoost feature selection and secondary decomposition *Renew. Energy* **203** 802–27

Kaiser R, Le Tertre A, Schwartz J, Gotway C A, Daley W R and Rubin C H 2007 The effect of the 1995 heat wave in Chicago on all-cause and cause-specific mortality *Am. J. Public Health* **97** S158–62
Karpinska L, Śmiech S, Gouveia J P and Palma P 2021 Mapping regional vulnerability to energy poverty in Poland *Sustainability* **13** 10694
Kephart J L, Sánchez B N, Moore J, Schinasi L H, Bakhtsiyarava M, Ju Y and Rodríguez D A 2022 City-level impact of extreme temperatures and mortality in Latin America *Nat. Med.* **28** 1700–5
Klinenberg E 1999 Denaturalizing disaster: a social autopsy of the 1995 Chicago heat wave *Theory Soc.* **28** 239–95
Klinenberg E 2022 *Heat Wave: A Social Autopsy of Disaster in Chicago* (University of Chicago Press)
Li F, Yigitcanlar T, Nepal M, Nguyen K, Dur F and Li W 2025 Mapping heat vulnerability in Australian capital cities: a machine learning and multi-source data analysis *Sustain. Cities Soc.* **119** 106079
Lim Y H, Bell M L, Kan H, Honda Y, Guo Y L L and Kim H 2015 Economic status and temperature-related mortality in Asia *Int. J. Biometeorol.* **59** 1405–12
Lounici K 2008 Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators (https://doi.org/10.1214/08-EJS177)
Lüthi S *et al* 2023 Rapid increase in the risk of heat-related mortality *Nat. Commun.* **14** 4894
Madrigano J, Mittleman M A, Baccarelli A, Goldberg R, Melly S, Von Klot S and Schwartz J 2013 Temperature, myocardial infarction, and mortality: effect modification by individual-and area-level characteristics *Epidemiology* **24** 439–46
Mallen E, Stone B and Lanza K 2019 A methodological assessment of extreme heat mortality modeling and heat vulnerability mapping in Dallas *Texas Urban Clim.* **30** 100528
Manson S, Schroeder J, Van Riper D, Knowles K, Kugler T, Roberts F and Ruggles S 2024 *IPUMS National Historical Geographic Information System: Version 19.0* (IPUMS)
Manware M, Dubrow R, Carrión D, Ma Y and Chen K 2022 Residential and race/ethnicity disparities in heat vulnerability in the United States *GeoHealth* **6** e2022GH000695
Muthukrishnan R and Rohini R L A S S O 2016 LASSO: a feature selection technique in predictive modeling for machine learning *2016 IEEE Int. Conf. on Advances in Computer Applications (ICACA)* (IEEE) pp 18–20
National Weather Service, NOAA 2022 The heat index equation (Weather Prediction Center) (available at: www.wpc.ncep.noaa.gov/html/heatindex_equation.shtml) (Accessed 24 May 2025)
Naughton M P, Henderson A, Mirabelli M C, Kaiser R, Wilhelm J L, Kieszak S M and McGeehin M A 2002 Heat-related mortality during a 1999 heat wave in Chicago *Am. J. Prev. Med.* **22** 221–7
Nguyen T T, Huang J Z and Nguyen T T 2015 Unbiased feature selection in learning random forests for high-dimensional data *Sci. World J.* **2015** 471371
Niu Y, Li Z, Gao Y, Liu X, Xu L, Vardoulakis S, Yue Y, Wang J and Liu Q 2021 A systematic review of the development and validation of the heat vulnerability index: major factors, methods, and spatial units *Curr. Clim. Change Rep.* **7** 87–97
O'Neill M S, Zanobetti A and Schwartz J 2003 Modifiers of the temperature and mortality association in seven US cities *Am. J. Epidemiol.* **157** 1074–82
O'Neill M S, Zanobetti A and Schwartz J 2005 Disparities by race in heat-related mortality in four US cities: the role of air conditioning prevalence *J. Urban Health.* **82** 191–7
Ostro B, Rauch S, Green R, Malig B and Basu R 2010 The effects of temperature and use of air conditioning on hospitalizations *Am. J. Epidemiol.* **172** 1053–61
Palecki M A, Changnon S A and Kunkel K E 2001 The nature and impacts of the July 1999 heat wave in the Midwestern United States: learning from the lessons of 1995 *Bull. Am. Meteorol. Soc.* **82** 1353–68
Pillai S K *et al* 2014 Heat Illness: Predictors of Hospital Admissions Among Emergency Department Visits—Georgia, 2002–2008 J Community Health **39** 90–8
Potash E, Brew J, Loewi A, Majumdar S, Reece A, Walsh J and Ghani R 2015 Predictive modeling for public health: preventing childhood lead poisoning *Proc. 21st ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining* pp 2039–47
Potash E, Ghani R, Walsh J, Jorgensen E, Lohff C, Prachand N and Mansour R 2020 Validation of a machine learning model to predict childhood lead poisoning *JAMA Netw. Open* **3** e2012734
Reid C E, O'neill M S, Gronlund C J, Brines S J, Brown D G, Diez-Roux A V and Schwartz J 2009 Mapping community determinants of heat vulnerability *Environ. Health Perspect.* **117** 1730–6
Sadilek A, Caty S, DiPrete L, Mansour R, Schenk T Jr, Bergtholdt M and Gabrilovich E 2018 Machine-learned epidemiology: real-time detection of foodborne illness at scale *npj Digit. Med.* **1** 36
Sarofim M C *et al* 2016 Chapter 2: temperature-related death and illness *USGCRP (U.S. Global Change Research Program), the Impacts of Climate Change on Human Health in the United States: A Scientific Assessment* pp 43–69
Schwingshackl C, Sillmann J, Vicedo-Cabrera A M, Sandstad M and Aunan K 2021 Heat stress indicators in CMIP6: estimating future trends and exceedances of impact-relevant thresholds *Earth Future* **9** e2020EF001885
Semenza J C, McCullough J E, Flanders W D, McGeehin M A and Lumpkin J R 1999 Excess hospital admissions during the July 1995 heat wave in Chicago *Am. J. Prev. Med.* **16** 269–77
Semenza J C, Rubin C H, Falter K H, Selanikio J D, Flanders W D, Howe H L and Wilhelm J L 1996 Heat-related deaths during the July 1995 heat wave in Chicago *New Engl. J. Med.* **335** 84–90
Shannon E Michael, Chen L, Yuan A, Chary A, Gabrielian S, Eisenman D P and Washington D L 2025 Extreme Heat, Social Factors, and Mortality Among California Veterans With Cardiometabolic Disease *JAMA Netw. Open* **8** e2545524
Shi X, Wong Y D, Li M Z F, Palanisamy C and Chai C 2019 A feature learning approach based on XGBoost for driving assessment and risk prediction *Accid. Anal. Prev.* **129** 170–9
Spangler K R, Weinberger K R and Wellenius G A 2019 Suitability of gridded climate datasets for use in environmental epidemiology *J. Expo. Sci. Environ. Epidemiol.* **29** 777–89
Speiser J L, Miller M E, Tooze J and Ip E 2019 A comparison of random forest variable selection methods for classification prediction modeling *Expert Syst. Appl.* **134** 93–101
Stang A, Standl F, Kowall B, Brune B, Böttcher J, Brinkmann M and Jöckel K H 2020 Excess mortality due to COVID-19 in Germany *J. Infect.* **81** 797–801
Sun S, Wang H, Cheng C, Chang Z and Huang D 2017 PCA-based heart sound feature generation for a ventricular septal defect discrimination *2017 14th Int. Computer Conf. on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)* (IEEE) pp 128–33
Tate E 2012 Social vulnerability indices: a comparative assessment using uncertainty and sensitivity analysis *Nat. Hazards.* **63** 325–47

Theng D and Bhoyar K K 2024 Feature selection techniques for machine learning: a survey of more than two decades of research *Knowl. Inf. Syst.* **66** 1575–637

Thornton M M, Shrestha R, Wei Y, Thornton P E and Kao S-C 2022 *Daymet: Daily Surface Weather Data on a 1-km Grid for North America, Version 4 R1* (https://doi.org/10.3334/ORNLDAAC/2129)

Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc. B* **58** 267–88

Toloo G S, Guo Y, Turner L, Qi X, Aitken P and Tong S 2014 Socio-demographic vulnerability to heatwave impacts in Brisbane, Australia: a time series analysis *Aust. N. Z. J. Public Health* **38** 430–5

U.S. Global Change Research Program (USGCRP) 2023 *Fifth National Climate Assessment* ed A R Crimmins, C W Avery, D R Easterling, K E Kunkel, B C Stewart and T K Maycock (U.S. Global Change Research Program) (available at: https://nca2023.globalchange.gov/)

Walton D and Hall A 2018 An assessment of high-resolution gridded temperature datasets over California *J. Clim.* **31** 3789–810

Wang S, Cai W, Tao Y, Sun Q C, Wong P P Y, Huang X and Liu Y 2023 Unpacking the inter- and intra-urban differences of the association between health and exposure to heat and air quality in Australia using global and local machine learning models *Sci. Total Environ.* **871** 162005

Whitman S, Good G, Donoghue E R, Benbow N, Shou W and Mou S 1997 Mortality in Chicago attributed to the July 1995 heat wave *Am. J. Public Health* **87** 1515–8

Wolf T, McGregor G and Analitis A 2014 Performance assessment of a heat wave vulnerability index for greater London, United Kingdom *Weather Clim. Soc.* **6** 32–46

World Health Organization 2024 Climate change, heat and health (available at: www.who.int/news-room/fact-sheets/detail/climate-change-heat-and-health)

World Meteorological Organization 2023 Climate change and heatwaves (available at: https://wmo.int/content/climate-change-and-heatwaves)

Zanobetti A, O'Neill M S, Gronlund C J and Schwartz J D 2013 Susceptibility to mortality in weather extremes: effect modification by personal and small-area characteristics *Epidemiology* **24** 809–19

Zhao L, Qing S, Bai J, Hao H, Li H, Shi Y and Yang R 2023 A hybrid optimized model for predicting evapotranspiration in early and late rice based on a categorical regression tree combination of key influencing factors *Comput. Electron. Agric.* **211** 108031

Zheng H, Yuan J and Chen L 2017 Short-term load forecasting using EMD-LSTM neural networks with a XGBoost algorithm for feature importance evaluation *Energies* **10** 1168

Zhou H, Ma L, Niu X, Xiang Y, Chen J, Su Y and Wu Q 2024 A novel hybrid model combined with ensemble embedded feature selection method for estimating reference evapotranspiration in the North China Plain *Agric. Water Manag.* **296** 108807