Article

Not peer-reviewed version

Atmospheric Exposures and Cardiovascular Mortality in U.S. Counties: Formaldehyde and Wet-Bulb Temperature as Leading Predictors

Samyak Shrestha , David J. Lary * , Shisir Ruwali , Faiz Ahmad

Posted Date: 23 March 2026

doi: 10.20944/preprints202603.1800.v1

Keywords: cardiovascular disease; CVD mortality; formaldehyde; air pollution; machine learning; XGBoost; SHAP; remote sensing; environmental health; county-level; United States

Article

Atmospheric Exposures and Cardiovascular Mortality in U.S. Counties: Formaldehyde and Wet-Bulb Temperature as Leading Predictors

Samyak Shrestha, David J. Lary *, Shisir Ruwali and Faiz Ahmad

Department of Physics, The University of Texas at Dallas, Richardson, TX 75080, USA
* Correspondence: david.lary@utdallas.edu

Abstract

Cardiovascular disease (CVD) is the leading cause of mortality in the United States, yet the role of atmospheric exposures as independent predictors of county-level CVD mortality remains poorly characterized. We integrated satellite-derived atmospheric data alongside socioeconomic, demographic, and livestock predictors across 24,487 county-year observations in the contiguous United States (2012–2019) and applied an XGBoost model with SHAP-based interpretability to identify the leading predictors of county-level CVD mortality (Test R^2 = 0.712, RMSE = 29.26 per 100,000). Three of the top ten predictors were atmospheric variables. Ambient formaldehyde exposure frequency ranked second among all 43 predictors, exceeded only by educational attainment and surpassing poverty rate. Wet-bulb temperature, which captures the combined physiological burden of heat and humidity, ranked third. Sulphate aerosol mixing ratio ranked seventh. These atmospheric predictors contributed independently of socioeconomic covariates, indicating that atmospheric exposure carries information about county-level CVD mortality risk not captured by income or education alone. Integrating atmospheric exposure monitoring into county-level CVD surveillance alongside socioeconomic indicators may improve the identification of high-risk geographies.

Keywords: cardiovascular disease; CVD mortality; formaldehyde; air pollution; machine learning; XGBoost; SHAP; remote sensing; environmental health; county-level; United States

---

1. Introduction

Cardiovascular disease (CVD) is the leading cause of mortality in the United States, responsible for 941,652 deaths in 2022, surpassing all forms of cancer and unintentional injuries combined [1]. Despite substantial declines in age-standardized CVD mortality rates over the past four decades, progress has been unevenly distributed across the country. County-level analyses reveal persistent geographic disparities in CVD burden, with the highest mortality concentration extending from southeastern Oklahoma along the Mississippi River Valley to eastern Kentucky [2]. As shown in Figure 1, county-level CVD mortality rates in 2019 ranged from approximately 73 to 639 deaths per 100,000 population. These patterns suggest that CVD mortality reflects not only individual clinical risk factors but also place-based socioeconomic and environmental conditions that remain incompletely characterized.

 doi:10.20944/preprints202603.1800.v1

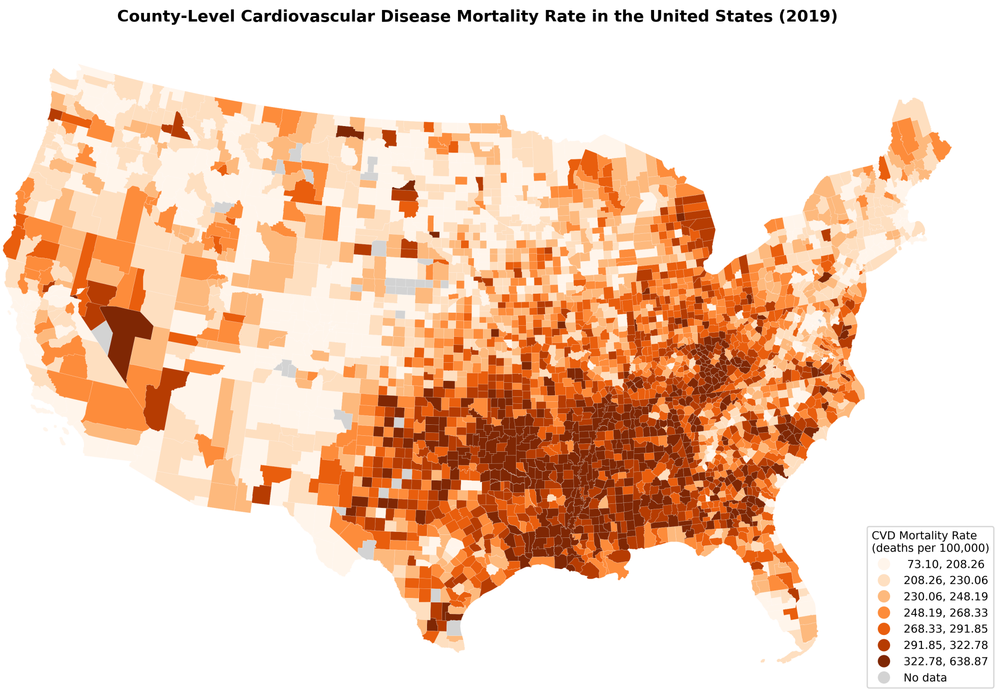

Figure 1. County-level age-standardized cardiovascular disease (CVD) mortality rate in the United States (2019). Values range from approximately 73 to 639 deaths per 100,000 population, with the highest burden concentrated in the Southeast, Appalachia, and the Mississippi River Delta. Data source: Institute for Health Metrics and Evaluation (IHME). Grey counties indicate missing data.

The socioeconomic determinants of CVD mortality are well established. Lower socioeconomic position is associated with a greater prevalence of cardiovascular risk factors and higher CVD incidence and mortality [3], with low socioeconomic status conferring a cardiovascular risk comparable to traditional clinical risk factors [4]. Socioeconomic disadvantage is also associated with higher rates of smoking, physical inactivity, and poor diet, each of which amplifies cardiovascular risk [5]. Higher educational attainment, in turn, is linked to better employment prospects, greater health literacy, and improved access to preventive care [6]. Despite the predictive strength of these socioeconomic factors, they do not fully account for the geographic variation in CVD mortality observed across US counties, pointing to a role for environmental exposures that operate independently of socioeconomic conditions.

Atmospheric exposures have emerged as an independent class of cardiovascular disease determinants. Air pollutants, including fine particulate matter ($PM_{2.5}$) and nitrogen dioxide, promote oxidative stress, immune cell activation, and vascular endothelial dysfunction, all established pathways to cardiovascular disease [7]. Long-term $PM_{2.5}$ exposure has been associated with an 8.8% increase in cardiovascular disease mortality per 10 $\mu g/m^3$ increase in concentration, based on a study of 53 million US Medicare beneficiaries, with no evidence of a lower safe threshold [8]. Formaldehyde, a ubiquitous indoor and ambient air pollutant, has received comparatively little attention in cardiovascular research. Like $PM_{2.5}$, it promotes oxidative stress through reactive oxygen species generation and glutathione depletion [9]. Experimental and epidemiological evidence further shows that formaldehyde exposure induces arrhythmia, myocardial infarction, heart failure, and atherosclerosis [10], and that ambient concentrations are associated with short-term increases in circulatory disease mortality [11]. Extreme heat and humidity, measured by wet-bulb temperature, also contribute to cardiovascular morbidity and excess mortality during heat events [7,12]. Ground-based monitoring of ambient formaldehyde is spatially sparse across the United States [13,14]. Satellite-derived atmospheric reanalysis products fill this gap, enabling county-level exposure characterization at national scale.

Despite growing evidence for both socioeconomic and atmospheric contributors to CVD mortality, no prior study has integrated these exposure classes within a unified predictive framework at the US

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2026 doi:10.20944/preprints202603.1800.v1

county level. We address this gap using an XGBoost model [15] with SHAP-based interpretability [16] and Bayesian hyperparameter optimization [17], applied to 24,487 county-year observations spanning 2012 to 2019. Our dataset integrates CVD mortality from the Institute for Health Metrics and Evaluation [18], socioeconomic data from the US Census American Community Survey [19], satellite-derived atmospheric and meteorological data from the Copernicus Atmosphere Monitoring Service and ERA5 reanalysis [20], and livestock density from the Food and Agriculture Organization Gridded Livestock of the World [21]. Among 43 predictors, educational attainment, ambient formaldehyde, and wet-bulb temperature emerge as the three leading drivers of county-level CVD mortality, with implications for environmental surveillance and cardiovascular public health.

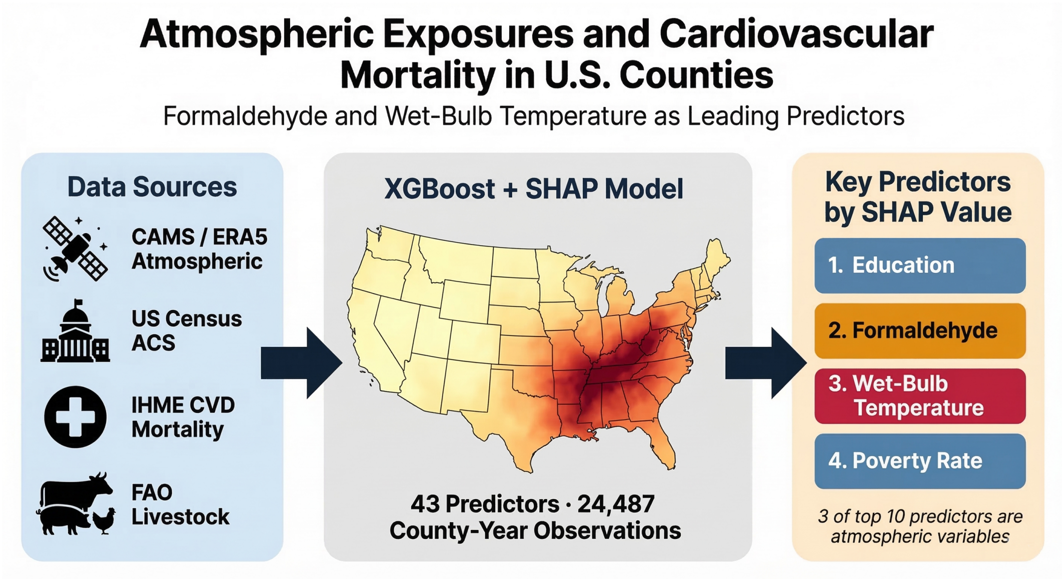

Figure 2. Study workflow. Four data sources are integrated into an XGBoost model with SHAP interpretability to identify the leading predictors of county-level CVD mortality across the contiguous United States.

2. Methodology

This study constructs a county-year panel dataset spanning 3,061 counties in the contiguous United States from 2012 to 2019. Four data sources were integrated to build the predictor and outcome variables: IHME cause-of-death estimates provided the target variable, the American Community Survey (ACS) contributed socioeconomic and demographic predictors, the Copernicus Atmosphere Monitoring Service (CAMS) and ERA5 reanalysis supplied atmospheric and meteorological predictors, and the Food and Agriculture Organization Gridded Livestock of the World (FAO GLW) provided livestock density features. An XGBoost regression model was trained with hyperparameters selected via Bayesian optimization. Model performance is reported as root mean square error (RMSE, in deaths per 100,000 population) and the coefficient of determination (R^2). Predictor contributions are quantified and ranked using SHAP values.

2.1. IHME Dataset

The target variable is the age-standardized cardiovascular disease (CVD) mortality rate, expressed as deaths per 100,000 population. Counties with older populations naturally have higher CVD mortality. Age standardization adjusts for this, producing rates that are directly comparable across counties regardless of their age structure. Estimates were drawn from the IHME Cause of Death dataset (released June 2023), which integrates county-level vital registration records and population data to produce annual, cause-specific mortality estimates for the contiguous United States [18]. The records were filtered to retain estimates for the total population across all racial groups and age-standardized mortality rates at the county level. The raw mortality proportion was converted to deaths per 100,000 population for numerical stability and to conform with standard epidemiological reporting

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2026 doi:10.20944/preprints202603.1800.v1

conventions. Data from 2012 to 2019 were selected to align temporally with the available predictor data from the ACS and CAMS sources.

2.2. American Community Survey Dataset

Socioeconomic and demographic predictors were obtained from the American Community Survey (ACS) five-year estimates, published by the US Census Bureau [19]. Five-year pooled estimates were used rather than one-year estimates because many counties, particularly rural ones, have populations too small for stable annual estimates, and for some smaller counties, one-year ACS data are not published at all. Pooling five years of survey data reduces margins of error and ensures full county coverage. The variables were retrieved via the Census API and covered poverty, income, educational attainment, unemployment, racial and ethnic composition, disability, and household characteristics; the complete list is provided in Appendix A. A total of 19 variables were initially retrieved; 10 were retained as predictors following the feature reduction procedure described in Section 2.5. Together, they cover the key socioeconomic determinants of CVD risk: income, education, poverty, and demographic composition.

2.3. CAMS and ERA5 Atmospheric Dataset

Atmospheric and meteorological predictor variables were derived from the CAMS EAC4 global reanalysis and the ERA5 reanalysis, both produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) [20]. CAMS EAC4 integrates satellite retrievals from the Ozone Monitoring Instrument (OMI), the Moderate Resolution Imaging Spectroradiometer (MODIS), and the Global Ozone Monitoring Experiment-2 (GOME-2) with atmospheric model assimilation to produce spatially continuous retrospective estimates of atmospheric composition at $0.75° \times 0.75°$ horizontal resolution. ERA5 contributes relative humidity at 0.25° resolution, as this variable is not available in CAMS EAC4. Three-hourly model output was averaged to produce annual mean concentrations for each variable. Fraction-of-Time (FoT) metrics were also computed for selected variables, defined as the proportion of three-hourly readings in a given year that exceeded the 75th percentile of that variable's long-term distribution. Annual values at county locations were obtained by multilinear interpolation of grid-point output to county boundary sample points, then averaged across sample points to produce a single county-level annual estimate. A total of 46 candidate atmospheric variables were analyzed. Of these, 26 were retained following the feature reduction procedure described in Section 2.5.

2.4. Livestock Dataset

Livestock density data were obtained from the Food and Agriculture Organization Gridded Livestock of the World (FAO GLW) dataset [21], which provides gridded population estimates for seven species at approximately 10 km spatial resolution: cattle, pigs, chickens, horses, sheep, goats, and ducks. County-level densities were computed using area-weighted zonal statistics applied to a 2019 National Land Cover Database (NLCD) land-use mask. GLW estimates are available at approximately five-year intervals (2010, 2015, and 2020); values for the intervening years from 2012 to 2019 were obtained by linear interpolation. Livestock operations are a recognized source of ammonia (NH_3), methane (CH_4), and nitrous oxide (N_2O) emissions, which contribute to regional atmospheric degradation and may elevate ambient concentrations of secondary pollutants associated with cardiovascular disease burden [22]. Seven livestock density features were retained in the final dataset.

 doi:10.20944/preprints202603.1800.v1

Table 1. Summary of data sources used to construct the county-year panel dataset (2012–2019). Feature counts for ACS, CAMS/ERA5, and FAO GLW reflect predictors retained after feature reduction.

Source	Description	Features (N)
IHME	Age-standardized CVD mortality rate (deaths per 100,000)	1 (target)
ACS 5-year estimates	Socioeconomic and demographic predictors	10
CAMS/ERA5	Atmospheric and meteorological predictors	26
FAO GLW	Livestock density by species	7
Predictors total		**43**

2.5. Data Processing

The final integrated dataset comprises 24,487 county-year observations spanning 3,061 US counties from 2012 to 2019, assembled by merging the four source datasets on county FIPS code and year. Records with missing values were excluded via listwise deletion. Multicollinearity was reduced by removing one feature from each correlated pair with $|r| > 0.85$, with the feature to drop determined by hierarchical clustering of the full correlation matrix. This reduced the initial candidate predictor pool to the 43 features retained in the final model. The dataset was partitioned into training and test sets using GroupShuffleSplit with a test fraction of 0.20 and county FIPS code as the grouping variable, so that no county appears in both partitions. This ensures the model is evaluated on counties not seen during training. The resulting partition yields 19,583 training observations across 2,448 counties and 4,904 test observations across 613 counties, with no county overlap between sets.

2.6. Modeling Approach

The modeling objective was to predict county-level age-standardized CVD mortality rates from the 43-feature predictor set. An XGBoost regression model [15] was trained as the primary estimator. Hyperparameters were selected using BayesSearchCV from the scikit-optimize library [17], which uses a Gaussian process surrogate model to explore the hyperparameter space more efficiently than grid or random search. The search ran for 30 iterations with 5-fold GroupKFold cross-validation, using county FIPS as the grouping variable to prevent data leakage across folds. Model selection was based on R^2 scoring. Table 2 reports the search bounds and the optimal values selected by the procedure. Following hyperparameter selection, the model was refit on the full training set and evaluated on the held-out test set. RMSE is expressed in deaths per 100,000 population.

Table 2. XGBoost hyperparameter search bounds and optimal values selected by BayesSearchCV. Log-uniform distributions were used for `learning_rate`, `reg_alpha`, and `reg_lambda`.

Parameter	Search Range	Optimal Value
`n_estimators`	150–800	800
`max_depth`	3–7	5
`learning_rate`	0.03–0.25 (log-uniform)	0.03
`subsample`	0.60–0.90	0.7042
`colsample_bytree`	0.50–0.85	0.7814
`reg_alpha`	0.05–8.0 (log-uniform)	0.05
`reg_lambda`	0.50–8.0 (log-uniform)	8.00
`min_child_weight`	3–15	3
Best CV R^2		0.7102

2.7. Modeling Interpretability

Three complementary methods were applied to interpret the trained model and assess predictor importance. First, permutation importance was computed on the test set by randomly shuffling each predictor in turn and measuring the resulting drop in R^2. Second, SHAP values were computed for all test-set predictions using TreeExplainer [16], which calculates exact Shapley values for tree-

 doi:10.20944/preprints202603.1800.v1

based models. SHAP values quantify each feature's positive or negative contribution to individual predictions. Dependence plots were generated for the three highest-ranked predictors to show how their effect on predicted CVD mortality varies across their value range. Third, a feature ablation study was conducted by training separate models on the top 20, top 10, and top 5 predictors ranked by SHAP importance and comparing RMSE and R^2 across the reduced feature sets to assess how much predictive performance depends on the full predictor set. Residual analysis was performed on the top-20 feature model to examine whether prediction errors were randomly distributed or showed systematic patterns by geography or CVD mortality level.

3. Results

3.1. Model Performance

The full 43-feature model achieved a test R^2 of 0.712 and a test RMSE of 29.26 deaths per 100,000, with a mean absolute error of 22.03 per 100,000 on the held-out test set of 4,904 county-year observations. The adjusted R^2 on the test set was 0.709, confirming that the result is not inflated by the number of predictors. On the training set, the model achieved R^2 = 0.869 and RMSE = 19.86 per 100,000. The gap between training and test R^2 (0.157) follows from the evaluation design. The test set comprises counties entirely withheld from training, so the model must generalize to regions it has never seen rather than interpolate across years within familiar counties. Predicted versus observed CVD mortality rates are shown in Figure 3.

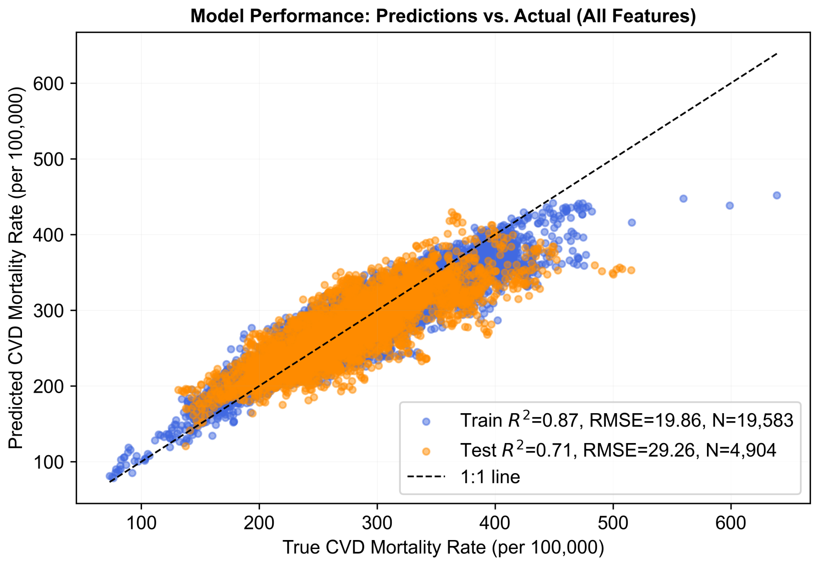

Figure 3. Predicted versus observed age-standardized CVD mortality rates (deaths per 100,000) on the held-out test set (n = 4,904 county-year observations). The dashed line indicates perfect agreement.

Residual diagnostics for the top-20 model are shown in Figure 4. Residuals were centered at zero across the full range of fitted values and showed no systematic relationship with poverty rate. The distribution of residuals was approximately normal. No directional bias was apparent at any level of predicted CVD mortality.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2026 doi:10.20944/preprints202603.1800.v1

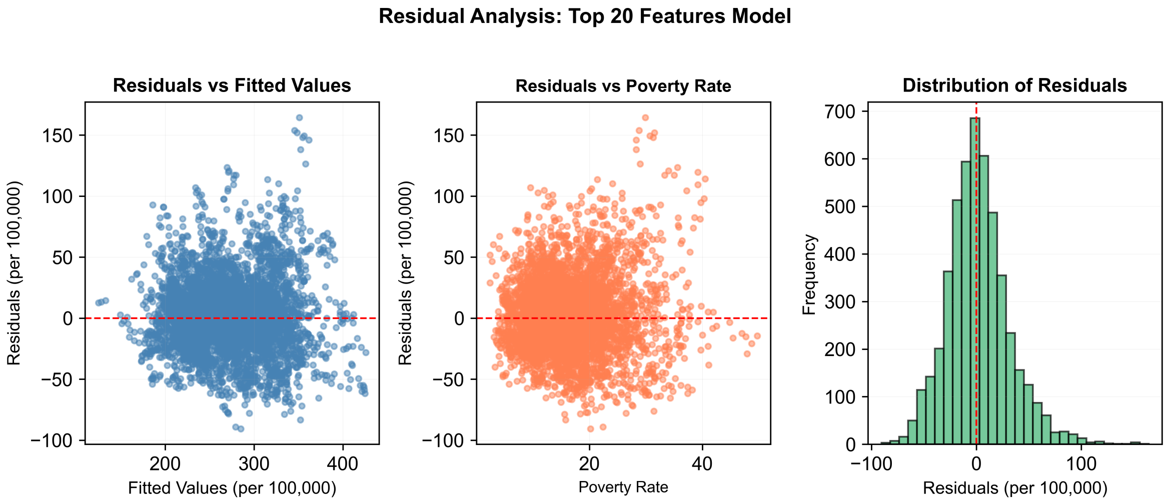

Figure 4. Residual diagnostics for the top-20 feature model. Left: residuals versus fitted values. Center: residuals versus poverty rate. Right: distribution of residuals. All residuals are in deaths per 100,000 population.

3.2. Feature Importance

SHAP values were computed for all test-set predictions to rank predictors by their contribution to predicted CVD mortality. The SHAP beeswarm plot is shown in Figure 5. Permutation importance, computed independently on the test set, yielded consistent rankings for the top predictors (Figure 6).

The percentage of adults with a bachelor's degree or higher (Bachelor's Degree or Higher %) was the strongest predictor, with higher values associated with lower predicted CVD mortality. FoT Formaldehyde Above the 75th Percentile ranked second and Wet-bulb Temperature third. Both showed positive associations with predicted CVD mortality, with higher exposure corresponding to higher predicted mortality rates. Poverty rate ranked fourth, with higher rates associated with higher predicted mortality. Among the top ten predictors, four were drawn from the CAMS/ERA5 atmospheric and meteorological dataset (FoT Formaldehyde, Wet-bulb Temperature, Sulphate Aerosol Mixing Ratio, and Leaf Area Index for High Vegetation), four were socioeconomic, and two were demographic.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2026 doi:10.20944/preprints202603.1800.v1

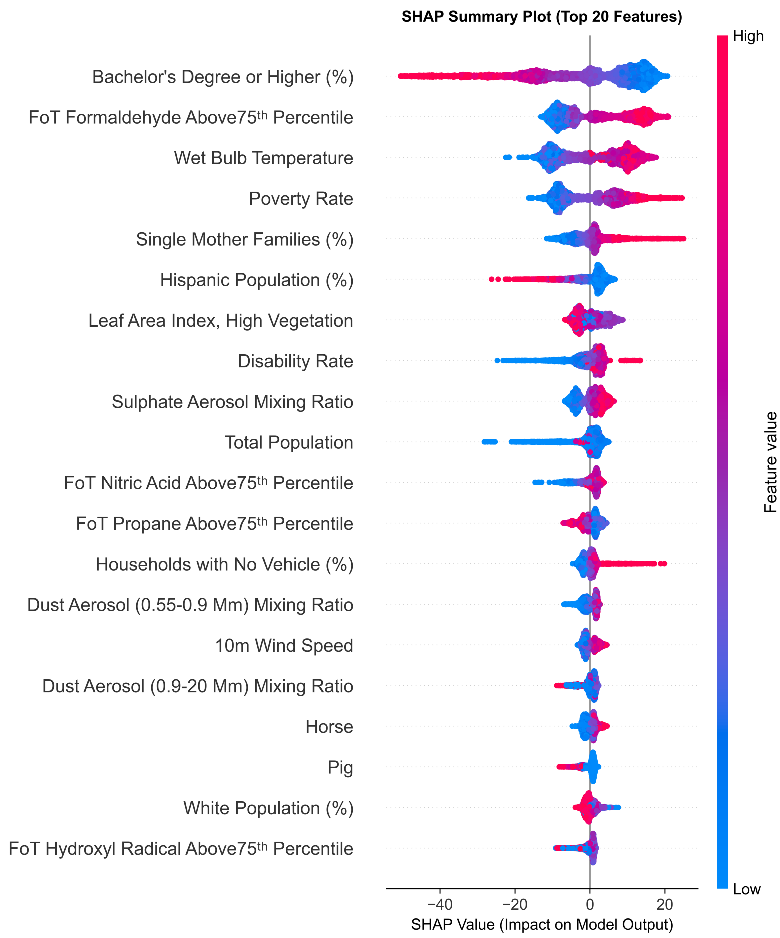

Figure 5. SHAP beeswarm plot for the full 43-feature model. Each row represents one predictor; each point represents one test-set observation. Point color indicates the feature value (red = high, blue = low). Horizontal position indicates the direction and magnitude of each feature's contribution to the predicted CVD mortality rate.

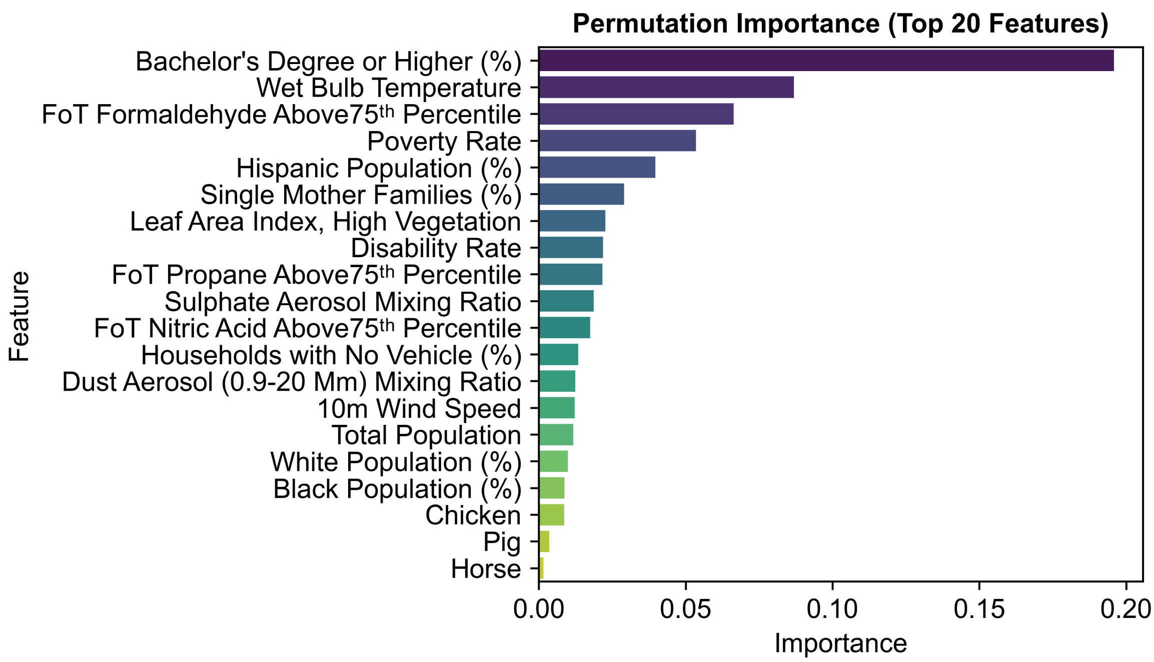

Figure 6. Permutation importance for the top-20 feature model. Each bar shows the mean decrease in test R^2 when the predictor's values are randomly shuffled. The ranking is consistent with the SHAP analysis.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 23 March 2026 doi:10.20944/preprints202603.1800.v1

SHAP dependence plots for the three highest-ranked predictors are shown in Figure 7. Educational attainment showed a monotonic negative relationship with predicted CVD mortality. FoT Formaldehyde Above the 75th Percentile showed a positive association that steepened at higher exposure levels, consistent with a non-linear dose-response relationship. Wet-bulb Temperature showed a positive association across its full observed range.

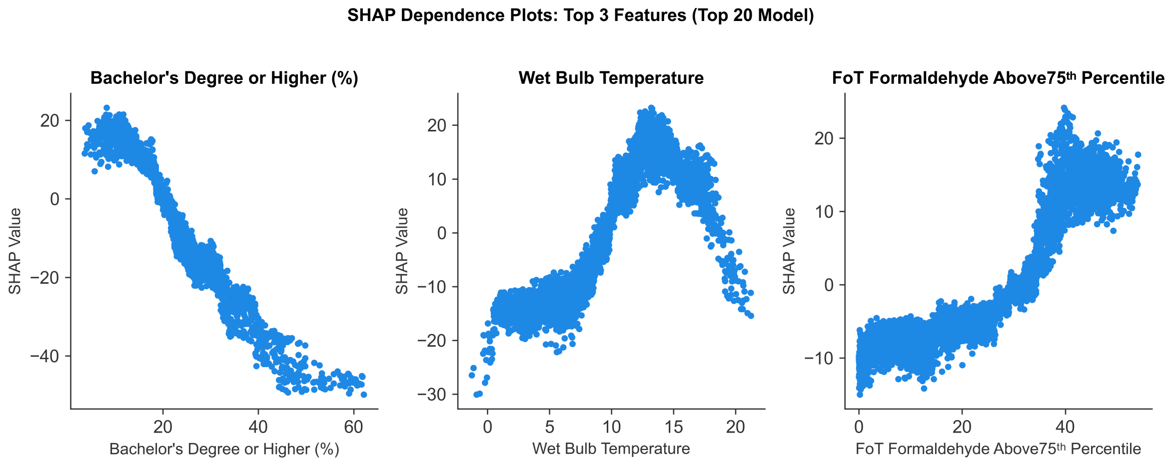

Figure 7. SHAP dependence plots for the three highest-ranked predictors: Bachelor's Degree or Higher % (top), FoT Formaldehyde Above the 75th Percentile (middle), and Wet-bulb Temperature (bottom). The x-axis shows the feature value; the y-axis shows the SHAP value (contribution to predicted CVD mortality). Point color encodes the value of the most interacting feature identified by SHAP.

3.3. Feature Ablation Study

Table 3 summarizes test-set performance across all four feature subsets. The top-20 model achieved a test R^2 of 0.703 and a test RMSE of 29.73 deaths per 100,000, a difference of 0.009 in R^2 from the full model while using fewer than half the predictors. The top-10 model reduced test R^2 to 0.688, a decline of 0.024 from the full model. The sharpest degradation occurred at the top-5 subset, where test R^2 fell to 0.619 and test RMSE rose to 33.65 per 100,000. Figure 8 shows that performance gains are concentrated in the transition from five to ten features, with diminishing returns above ten. The top 20 features provide a near-equivalent representation of the full model, while further reduction below ten features results in meaningful performance loss.

Table 3. Test-set performance across feature subsets. RMSE and MAE are expressed in deaths per 100,000 population.

Feature Set	N	Test R^2	Test Adj R^2	Test RMSE	Test MAE
All features	43	0.712	0.709	29.26	22.03
Top 20	20	0.703	0.701	29.73	22.38
Top 10	10	0.688	0.687	30.47	22.66
Top 5	5	0.619	0.619	33.65	25.49

 doi:10.20944/preprints202603.1800.v1

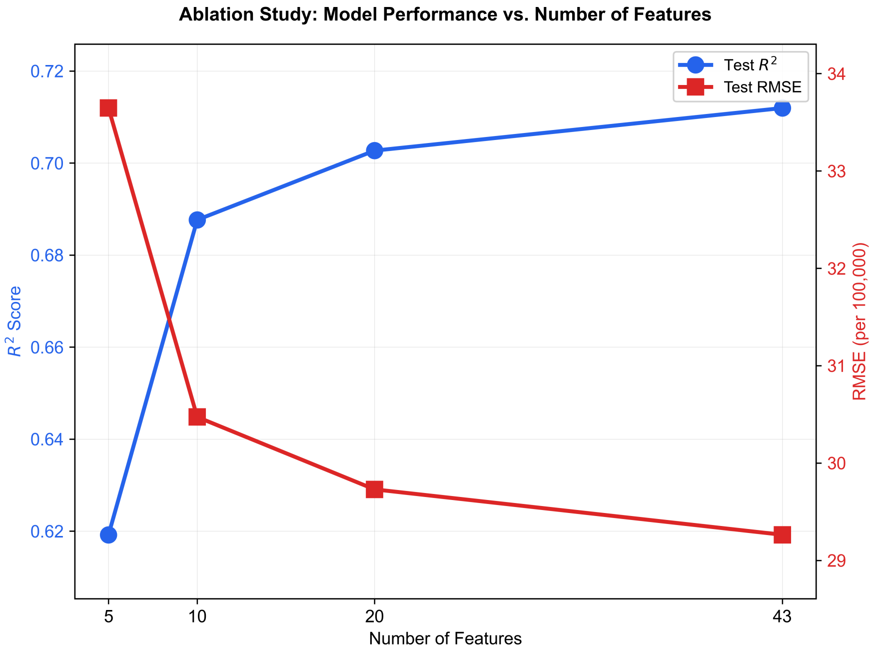

Figure 8. Ablation curve showing test R^2 (left axis) and test RMSE (right axis) as a function of the number of predictors retained. Each point corresponds to a model trained on the top *k* features ranked by SHAP importance.

4. Discussion

4.1. Atmospheric Exposures as Independent Predictors of CVD Mortality

FoT Formaldehyde Above the 75th Percentile ranked second, Wet-bulb Temperature ranked third, and Sulphate Aerosol Mixing Ratio ranked seventh among all 43 predictors by mean absolute SHAP value. All three are atmospheric variables. Each ranked above several socioeconomic predictors with well-established roles in cardiovascular epidemiology. Environmental exposures are recognized determinants of cardiovascular health through multiple biological pathways [7]. These atmospheric rankings held after the model accounted for socioeconomic predictors alongside them, so they reflect an independent signal rather than a proxy for poverty or low educational attainment.

FoT Formaldehyde Above the 75th Percentile measures the fraction of time during a year when county-level formaldehyde concentrations exceeded the 75th percentile of the full dataset. Formaldehyde has direct cardiovascular toxicity [10]. Experimental evidence documents its capacity to induce arrhythmia, myocardial infarction, heart failure, and atherosclerosis through oxidative stress, endothelial dysfunction, and vascular inflammation. At the population level, higher ambient formaldehyde concentrations have been associated with increased circulatory disease mortality across Chinese counties [11]. Counties with higher chronic formaldehyde exposure had higher predicted CVD mortality in this model, consistent with both the experimental and epidemiological evidence. Figure 9 shows the geographic distribution of FoT Formaldehyde across US counties, averaged over 2012–2019, with the highest exposure frequencies concentrated in the Southeast and Gulf Coast — regions that also bear a disproportionate share of cardiovascular disease mortality in the United States.

doi:10.20944/preprints202603.1800.v1

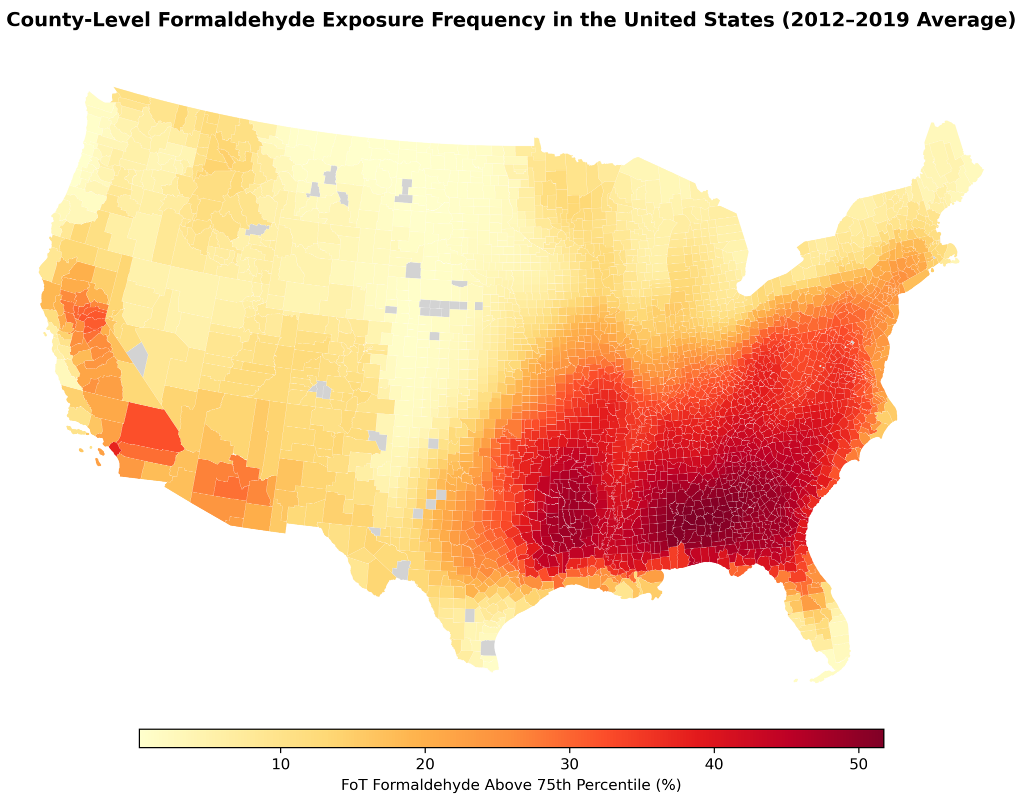

Figure 9. County-level mean FoT Formaldehyde Above the 75th Percentile across the contiguous United States, averaged over 2012–2019. Values represent the percentage of 3-hourly time steps in a year during which county-level formaldehyde concentrations exceeded the 75th percentile threshold of the full dataset.Grey counties indicate missing data.

Wet-bulb temperature ranked third. Unlike dry-bulb temperature, wet-bulb temperature captures both heat and humidity, reflecting conditions under which the body cannot effectively dissipate heat through evaporation [12]. Heat stress imposes a direct hemodynamic burden on the cardiovascular system. The body maintains core temperature through peripheral vasodilation and elevated cardiac output, creating demands that increase cardiovascular strain [7]. In individuals with pre-existing coronary artery disease or reduced cardiac reserve, this added demand can precipitate myocardial ischemia and arrhythmia. Large studies confirm that a substantial share of deaths linked to extreme temperatures involve cardiovascular causes [23] and that heat-related mortality is disproportionately cardiovascular in nature [24]. Counties with higher wet-bulb temperatures had higher predicted CVD mortality in this model, consistent with this evidence. Figure 10 shows the geographic distribution of mean wet-bulb temperature across US counties, with the highest values concentrated in the Gulf Coast and the Deep South.

 doi:10.20944/preprints202603.1800.v1

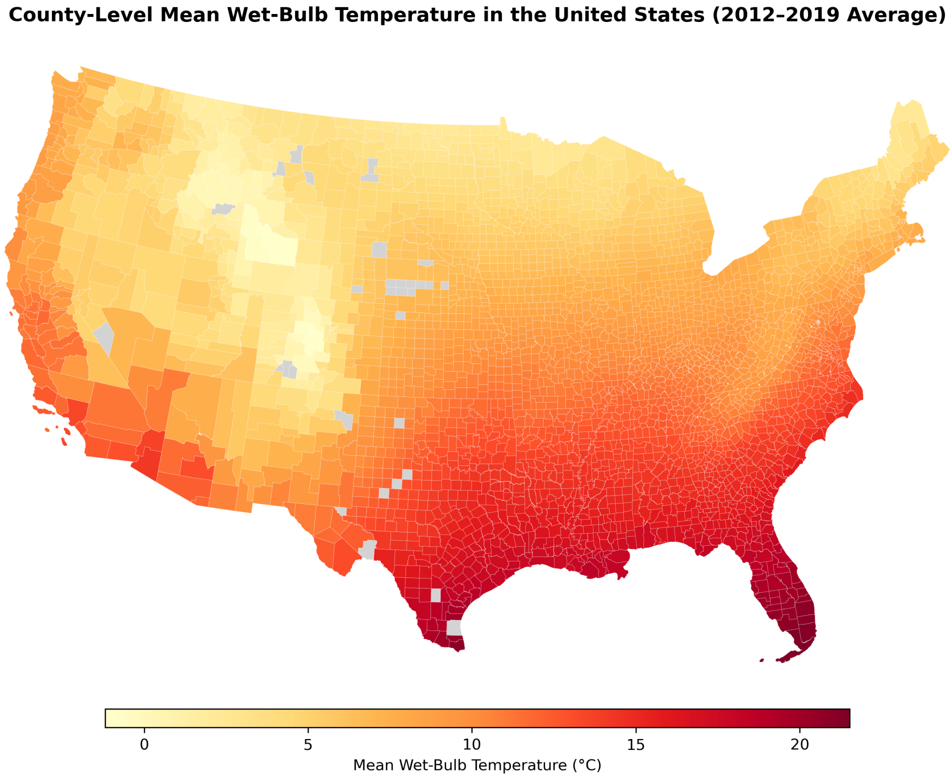

Figure 10. County-level mean wet-bulb temperature (°C) across the contiguous United States, averaged over 2012–2019. Wet-bulb temperature reflects the combined physiological burden of heat and humidity, representing conditions under which evaporative cooling is limited. Grey counties indicate missing data.

Sulphate aerosol mixing ratio ranked seventh. Sulphates are a major component of fine particulate matter ($PM_{2.5}$), and long-term $PM_{2.5}$ exposure has been linked to higher cardiovascular mortality in large US population studies [25,26]. Among 53 million US Medicare beneficiaries, $PM_{2.5}$ was associated with elevated cardiovascular cause-specific mortality [8]. That sulphate aerosol ranked among the top predictors independently of the other atmospheric variables in the model suggests that sulphates specifically, not air pollution in general, carry predictive signal for county-level CVD mortality.

Leaf Area Index for High Vegetation ranked eighth. This variable measures tree canopy density and likely captures geographic variation in land cover rather than a direct exposure pathway.

Taken together, three of the top ten SHAP predictors are atmospheric variables, each linked to the cardiovascular system through a distinct biological or epidemiological pathway. These findings suggest that atmospheric exposures carry independent predictive signal for county-level CVD mortality, beyond what socioeconomic predictors alone would provide.

4.2. Socioeconomic and Demographic Determinants

The percentage of adults with a bachelor's degree or higher (Bachelor's Degree or Higher %) ranked first among all 43 predictors, with higher county-level values associated with lower predicted CVD mortality. This is consistent with a large body of evidence linking population educational levels to cardiovascular outcomes [3,4]. At the county level, educational attainment correlates with income, occupational risk, healthcare access, and health behaviors, making it a compound indicator of socioeconomic conditions rather than a single pathway [6]. Poverty rate ranked fourth and showed a positive SHAP direction, consistent with the social determinants of health literature [27,28]. The two predictors are correlated but carry partially non-overlapping information, and the presence of both in the top five suggests each contributes independently to the variation in county-level CVD mortality.

The percentage of Hispanic residents (Hispanic Population %) ranked fifth and showed a negative SHAP direction. Counties with larger Hispanic populations had lower predicted CVD mortality after accounting for socioeconomic and atmospheric covariates. This is consistent with the Hispanic paradox, the empirically documented pattern in which Hispanic adults in the United States experience lower cardiovascular mortality than non-Hispanic White adults despite a less favorable socioeconomic profile [29]. The percentage of single mother families (Single Mother Families %) ranked sixth, with a positive SHAP direction, consistent with evidence that single-parent household status is associated with elevated cardiovascular risk [30]. Disability Rate ranked tenth and also showed a positive SHAP direction.

4.3. Strengths and Limitations

This study has several methodological strengths. Model evaluation used geographic train-test separation by county FIPS code, so that every county in the test set was entirely absent from training. This design ensures that no county contributed observations to both the training and test sets, which is the appropriate holdout for data structured as repeated observations across counties and years. SHAP values were computed using TreeExplainer, which provides exact Shapley values rather than approximations, enabling reliable attribution of model predictions to individual features across the full predictor set. The predictor space spans socioeconomic, demographic, atmospheric, and land-surface domains, reflecting the multifactorial nature of county-level CVD mortality. The feature ablation study confirmed that predictive performance is distributed across the predictor set and is not driven by any single variable or small subset.

Several limitations should be acknowledged. All associations are ecological. The data are aggregated at the county level, and individual-level causal inference is not possible from county-level patterns. Each county-year is treated as an independent observation. The model does not track how conditions within individual counties change from year to year. Whether the predictor-outcome associations documented here persist or change as environmental and socioeconomic conditions evolve cannot be determined from this analysis. SHAP values identify which predictors contribute most to model predictions, but they do not establish whether those predictors have causal effects on CVD mortality. A predictor that ranks highly may reflect a direct relationship with CVD mortality or may be correlated with an unmeasured variable that is the true driver. Unmeasured county-level factors, including healthcare access quality, diet patterns, physical activity rates, and smoking prevalence, may account for part of the residual variance. Educational attainment and poverty rate, both of which ranked among the top five predictors, are established correlates of healthcare access, health behaviors, and smoking prevalence, and may capture part of the variation attributable to these unmeasured factors [6,27]. Atmospheric data were drawn from CAMS/ERA5 reanalysis at 0.75° spatial resolution. This resolution is coarser than many US counties, especially in the densely populated eastern states, meaning that multiple counties may share identical atmospheric values. The resulting spatial smoothing likely attenuates the observed associations between atmospheric exposures and CVD mortality. The model also does not account for spatial autocorrelation. CVD mortality rates in neighboring counties tend to be more similar than those in distant counties, and ignoring this structure may understate the uncertainty of the results.

5. Conclusions

An XGBoost model trained on 43 county-level predictors explained 71.2% of the variance in cardiovascular disease mortality across US counties (Test R^2 = 0.712, RMSE = 29.26 per 100,000). Three of the top ten predictors by SHAP value were atmospheric variables. FoT Formaldehyde Above the 75th Percentile ranked second, Wet-bulb Temperature ranked third, and Sulphate Aerosol Mixing Ratio ranked seventh. Each ranked above several socioeconomic predictors with well-established roles in cardiovascular epidemiology. The atmospheric predictors contributed predictive value beyond what socioeconomic variables alone would provide.

Atmospheric exposure data may complement socioeconomic indicators in county-level CVD surveillance and improve the identification of high-risk geographies. Formaldehyde concentrations, wet-bulb temperature, and sulphate aerosol levels are routinely available from satellite and reanalysis products and could be incorporated into public health monitoring at low cost. Future work should

 doi:10.20944/preprints202603.1800.v1

examine the temporal dynamics of these associations and test whether finer-resolution atmospheric data improve predictive performance in densely populated regions.

Author Contributions: Conceptualization, S.S. and D.J.L.; methodology, S.S.; software, S.S.; validation, S.S.; formal analysis, S.S.; investigation, S.S.; resources, D.J.L.; data curation, S.S., S.R. and F.A.; writing—original draft preparation, S.S.; writing—review and editing, D.J.L.; visualization, S.S.; supervision, D.J.L.; project administration, D.J.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Code and processed data are available at https://github.com/samyakshrestha/predicting-cvd. Raw data sources are publicly available from the Institute for Health Metrics and Evaluation (IHME), U.S. Census Bureau American Community Survey, Copernicus Atmosphere Monitoring Service (CAMS), European Centre for Medium-Range Weather Forecasts (ECMWF) ERA5, and Food and Agriculture Organization (FAO) Gridded Livestock of the World as cited in this manuscript.

Acknowledgments: The authors used Claude (Anthropic) to assist with manuscript writing and coding. All outputs were reviewed, validated, and edited by the authors, who take full responsibility for the content.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Variable Descriptions

Table A1. Complete set of 43 predictor variables used in the XGBoost model, grouped by data source. FoT (Fraction of Time) denotes the percentage of 3-hourly time steps in a year during which county-level concentrations exceeded the 75th percentile threshold of the full dataset.

No.	Variable	Unit
Socioeconomic and Demographic Variables (American Community Survey, 5-year estimates)		
1	Poverty Rate	%
2	Bachelor's Degree or Higher (%)	%
3	Disability Rate	%
4	Total Population	count
5	Unemployment Rate	%
6	White Population (%)	%
7	Hispanic Population (%)	%
8	Black Population (%)	%
9	Households with No Vehicle (%)	%
10	Single Mother Families (%)	%
Atmospheric and Meteorological Variables (CAMS/ERA5)		
11	Land-sea Mask	–
12	Mean Sea Level Pressure	Pa
13	Dust Aerosol (0.55–0.9 μm) Mixing Ratio	$kgkg^{-1}$
14	Dust Aerosol (0.9–20 μm) Mixing Ratio	$kgkg^{-1}$
15	Hydrophilic Black Carbon Aerosol Mixing Ratio	$kgkg^{-1}$
16	Hydrophobic Black Carbon Aerosol Mixing Ratio	$kgkg^{-1}$
17	Hydrophobic Organic Matter Aerosol Mixing Ratio	$kgkg^{-1}$
18	Sea Salt Aerosol (0.5–5 μm) Mixing Ratio	$kgkg^{-1}$
19	Sea Salt Aerosol (5–20 μm) Mixing Ratio	$kgkg^{-1}$
20	Sulphate Aerosol Mixing Ratio	$kgkg^{-1}$
21	Leaf Area Index, High Vegetation	m^2m^{-2}
22	Leaf Area Index, Low Vegetation	m^2m^{-2}
23	Snow Depth	m w.e.
24	10m Wind Speed	ms^{-1}
25	Wet Bulb Temperature	K
26	FoT Carbon Monoxide Above 75th Percentile	%
27	FoT Ethane Above 75th Percentile	%
28	FoT Formaldehyde Above 75th Percentile	%

Table A1. *Cont.*

No.	Variable	Unit
Atmospheric and Meteorological Variables (CAMS/ERA5)		
29	FoT Hydroxyl Radical Above 75th Percentile	%
30	FoT Nitric Acid Above 75th Percentile	%
31	FoT Nitrogen Dioxide Above 75th Percentile	%
32	FoT Nitrogen Monoxide Above 75th Percentile	%
33	FoT Ozone Above 75th Percentile	%
34	FoT $PM_{2.5}$ Above 75th Percentile	%
35	FoT Propane Above 75th Percentile	%
36	FoT Sulphur Dioxide Above 75th Percentile	%
Livestock Density Variables (FAO Gridded Livestock of the World)		
37	Cattle	head km^{-2}
38	Chicken	head km^{-2}
39	Duck	head km^{-2}
40	Goat	head km^{-2}
41	Horse	head km^{-2}
42	Pig	head km^{-2}
43	Sheep	head km^{-2}

References

1. Martin, S.S.; Aday, A.W.; Allen, N.B.; Almarzooq, Z.I.; Anderson, C.A.M.; Arora, P.; Avery, C.L.; Baker-Smith, C.M.; Bansal, N.; Beaton, A.Z.; et al. 2025 Heart Disease and Stroke Statistics: A Report of US and Global Data From the American Heart Association. *Circulation* **2025**, *151*, e41–e660. Epub 2025 Jan 27, https://doi.org/10.1161/CIR.0000000000001303.
2. Roth, G.A.; Dwyer-Lindgren, L.; Bertozzi-Villa, A.; Stubbs, R.W.; Morozoff, C.; Naghavi, M.; Mokdad, A.H.; Murray, C.J.L. Trends and Patterns of Geographic Variation in Cardiovascular Mortality Among US Counties, 1980–2014. *JAMA* **2017**, *317*, 1976–1992. https://doi.org/10.1001/jama.2017.4150.
3. Havranek, E.P.; Mujahid, M.S.; Barr, D.A.; Blair, I.V.; Cohen, M.S.; Cruz-Flores, S.; Davey-Smith, G.; Dennison-Himmelfarb, C.R.; Lauer, M.S.; Lockwood, D.W.; et al. Social Determinants of Risk and Outcomes for Cardiovascular Disease: A Scientific Statement From the American Heart Association. *Circulation* **2015**, *132*, 873–898. https://doi.org/10.1161/CIR.0000000000000228.
4. Schultz, W.M.; Kelli, H.M.; Lisko, J.C.; Varghese, T.; Shen, J.; Sandesara, P.; Quyyumi, A.A.; Taylor, H.A.; Gulati, M.; Harold, J.G.; et al. Socioeconomic Status and Cardiovascular Outcomes: Challenges and Interventions. *Circulation* **2018**, *137*, 2166–2178. https://doi.org/10.1161/CIRCULATIONAHA.117.029652.
5. Pampel, F.C.; Krueger, P.M.; Denney, J.T. Socioeconomic Disparities in Health Behaviors. *Annual Review of Sociology* **2010**, *36*, 349–370. https://doi.org/10.1146/annurev.soc.012809.102529.
6. Zajacova, A.; Lawrence, E.M. The relationship between education and health: reducing disparities through a contextual approach. *Annual Review of Public Health* **2018**, *39*, 273–289. Final edited form published January 12, 2018, https://doi.org/10.1146/annurev-publhealth-031816-044628.
7. Blaustein, J.R.; Quisel, M.J.; Hamburg, N.M.; Wittkopp, S. Environmental Impacts on Cardiovascular Health and Biology: An Overview. *Circulation Research* **2024**, *134*, 1048–1060. https://doi.org/10.1161/CIRCRESAHA.123.323613.
8. Wang, B.; Eum, K.D.; Kazemiparkouhi, F.; Li, C.; Manjourides, J.; Pavlu, V.; Suh, H.H. The impact of long-term $PM_{2.5}$ exposure on specific causes of death: exposure-response curves and effect modification among 53 million U.S. Medicare beneficiaries. *Environmental Health* **2020**. https://doi.org/10.1186/s12940-020-00575-0.
9. Ghelli, F.; Bellisario, V.; Squillacioti, G.; Panizzolo, M.; Santovito, A.; Bono, R. Formaldehyde in Hospitals Induces Oxidative Stress: The Role of GSTT1 and GSTM1 Polymorphisms. *Toxics* **2021**, *9*, 178. https://doi.org/10.3390/toxics9080178.
10. Zhang, Y.; Yang, Y.; He, X.; Yang, P.; Zong, T.; Sun, P.; Sun, R.C.; Yu, T.; Jiang, Z. The cellular function and molecular mechanism of formaldehyde in cardiovascular disease and heart development. *Journal of Cellular and Molecular Medicine* **2021**, *25*, 5358–5371. https://doi.org/10.1111/jcmm.16602.
11. Ban, J.; Su, W.; Zhong, Y.; Liu, C.; Li, T. Ambient formaldehyde and mortality: A time series analysis in China. *Science Advances* **2022**, *8*, eabm4097. https://doi.org/10.1126/sciadv.abm4097.

12. Raymond, C.; Matthews, T.; Horton, R.M. The emergence of heat and humidity too severe for human tolerance. *Science Advances* **2020**, *6*, eaaw1838. https://doi.org/10.1126/sciadv.aaw1838.
13. Zhu, L.; Jacob, D.J.; Keutsch, F.N.; Mickley, L.J.; Scheffe, R.D.; Strum, M.; González Abad, G.; Chance, K.; Yang, K.; Rappenglück, B.; et al. Formaldehyde (HCHO) as a Hazardous Air Pollutant: Mapping Surface Air Concentrations from Satellite and Inferring Cancer Risks in the United States. *Environmental Science & Technology* **2017**, *51*, 5650–5657. https://doi.org/10.1021/acs.est.7b01356.
14. Wang, P.; Holloway, T.; Bindl, M.; Harkey, M.; De Smedt, I. Ambient Formaldehyde over the United States from Ground-Based (AQS) and Satellite (OMI) Observations. *Remote Sensing* **2022**, *14*, 2191. https://doi.org/10.3390/rs14092191.
15. Chen, T.; Guestrin, C. XGBoost: A scalable tree boosting system. In Proceedings of the Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016, pp. 785–794.
16. Lundberg, S.M.; Lee, S.I. A Unified Approach to Interpreting Model Predictions. *arXiv preprint arXiv:1705.07874* **2017**. Submitted 22 May 2017; revised 25 Nov 2017, https://doi.org/10.48550/arXiv.1705.07874.
17. Wild Tree Tech.; Google Brain.; University of Liège.; Saarland University. scikit-optimize: Sequential model-based optimization in Python. https://scikit-optimize.github.io/stable/modules/generated/skopt.BayesSearchCV.html, 2020. Version 0.8.1. DOI: 10.5281/zenodo.4014775.
18. for Health Metrics, I.; (IHME), E. United States Mortality Rates and Life Expectancy by County, Race, and Ethnicity 2000–2019. Global Health Data Exchange (GHDx), 2022. Accessed via GHDx record "United States Life Expectancy by County, Race & Ethnicity 2000-2019".
19. U.S. Census Bureau. American Community Survey 5-Year Data (2009–2023). https://www.census.gov/data/developers/data-sets/acs-5year.html, 2024. Accessed: 2026-01-14.
20. Inness, A.; Ades, M.; Agustí-Panareda, A.; Barré, J.; Benedictow, A.; Blechschmidt, A.M.; Dominguez, J.J.; Engelen, R.; Eskes, H.; Flemming, J.; et al. The CAMS reanalysis of atmospheric composition. *Atmospheric Chemistry and Physics* **2019**, *19*, 3515–3556. https://doi.org/10.5194/acp-19-3515-2019.
21. Gilbert, M.; Nicolas, G.; Cinardi, G.; Van Boeckel, T.P.; Vanwambeke, S.O.; Wint, G.W.; Robinson, T.P. Global distribution data for cattle, buffaloes, horses, sheep, goats, pigs, chickens and ducks in 2010. *Scientific Data* **2018**, *5*, 1–12. https://doi.org/10.1038/sdata.2018.227.
22. Anestis, V.; Umar, W.; Dragoni, F.; van der Weerden, T.J.; Hassouna, M.; Noble, A.; Bartzanas, T.; Amon, B. Mitigation of greenhouse gas and ammonia emissions due to livestock housing management practices: Analysis of the DATAMAN database. *Biosystems Engineering* **2025**, *258*, 104260. https://doi.org/10.1016/j.biosystemseng.2025.104260.
23. Zhao, Q.; Guo, Y.; Ye, T.; Gasparrini, A.; Tong, S.; Overcenco, A.; Urban, A.; Schneider, A.; Entezari, A.; Vicedo-Cabrera, A.M.; et al. Global, regional, and national burden of mortality associated with non-optimal ambient temperatures from 2000 to 2019: a three-stage modelling study. *The Lancet Planetary Health* **2021**, *5*, e415–e425. https://doi.org/10.1016/S2542-5196(21)00081-4.
24. Gallo, E.; Quijal-Zamorano, M.; Méndez Turrubiates, R.F.; Tonne, C.; Basagaña, X.; Achebak, H.; Ballester, J. Heat-related mortality in Europe during 2023 and the role of adaptation in protecting health. *Nature Medicine* **2024**, *30*, 3101–3105. https://doi.org/10.1038/s41591-024-03186-1.
25. Pope III, C.A.; Ezzati, M.; Dockery, D.W. Fine-Particulate Air Pollution and Life Expectancy in the United States. *New England Journal of Medicine* **2009**, *360*, 376–386. https://doi.org/10.1056/NEJMsa0805646.
26. Di, Q.; Wang, Y.; Zanobetti, A.; Wang, Y.; Koutrakis, P.; Choirat, C.; Dominici, F.; Schwartz, J.D. Air Pollution and Mortality in the Medicare Population. *New England Journal of Medicine* **2017**, *376*, 2513–2522. https://doi.org/10.1056/NEJMoa1702747.
27. Marmot, M. Social determinants of health inequalities. *The Lancet* **2005**, *365*, 1099–1104. https://doi.org/10.1016/S0140-6736(05)71146-6.
28. Singh, G.K.; Lee, H. Marked Disparities in Life Expectancy by Education, Poverty Level, Occupation, and Housing Tenure in the United States, 1997–2014. *Int. J. MCH AIDS* **2021**, *10*, 7–18. Epub 2020 Dec 30, https://doi.org/10.21106/ijma.402.

 doi:10.20944/preprints202603.1800.v1

29. Khan, S.U.; Lone, A.N.; Yedlapati, S.H.; Dani, S.S.; Khan, M.Z.; Watson, K.E.; Parwani, P.; Rodriguez, F.; Cainzos-Achirica, M.; Michos, E.D. Cardiovascular Disease Mortality Among Hispanic Versus Non-Hispanic White Adults in the United States, 1999 to 2018. *Journal of the American Heart Association* **2022**, *11*, e022857. https://doi.org/10.1161/JAHA.121.022857.
30. Borrowman, J.D.; Pageau, L.M.; Cameron, N.A.; Carnethon, M.R.; Fields, N.D. Cardiovascular Health of Single Mothers: A Systematic Review. *Journal of the American Heart Association* **2025**, *14*, e043296. https://doi.org/10.1161/JAHA.125.043296.