Investigating the incidence and potential risk factors for hypertension-related mortalities in adults in the United States.

Shaun van den Hurk

Warning: package 'magick' was built under R version 4.4.3

**Authors**

* First Author: Shaun van den Hurk

**Author affiliations**

1. College of Veterinary Medicine, University of Georgia, Athens, GA, USA.

These authors contributed equally to this work.

Corresponding author: shaun.vandenhurk@uga.edu

Disclaimer: The opinions expressed in this article are the author’s own and don’t reflect their employer.

# 1. Summary/Abstract

Lifestyle diseases such as cardiovascular disease make up the highest causes of death in developed nations such as the United States. This project focuses on hypertension-related cardiovascular disease and hopes to help provide a greater understanding of the dynamics and patterns that may exist with hypertension and cardiovascualr diseased related mortality in the United States. The project seeks to evaluate data from the CDC on the rates of hypertension-related cardiovascular disease mortality and the associated demographic factors for this collected data, including county, age group, race/ethnicity, and sex. The data represents information collected from 2000 to 2019. This will be used to see if there are any potential patterns or correlation between the rate of mortality and these other demographic factors within the data.

# 2. Introduction

## 2.1 General Background Information

Lifestyle diseases such as cardiovascular disease (CVD) represent some of the most significant causes of death globally, both in low-income regions and high-income countries such as the United States of America (USA). Cardiovascular diseases are a broad umbrella term including different clinical manifestations such as hypertension, stroke, and ischemic heart disease. Cardiovascular diseases were cited as the leading cause of death globally in 2022 (..1..), in the USA they were responsible for one in every five deaths nationally (…2.). Hypertension is commonly identified as a significant risk factor for multiple cardiovascular conditions, and one that is modifiable, and thus represents a good metric for potential public health interventions (…3.). However, not all demographics appear to be equally affected by hypertension and CVD, with differences in rates of CVD and hypertension in people from different races, sexes, age groups, geographic location, income levels, and other groups in previous studies (..4..). A deeper understanding of these factors helps us to get a deeper understanding of the multifactorial nature of CVD and also helps guide effective public health interventions. Additionally, it is important to understand if there have been any differences in these rates of CVD and hypertension over time, and if there are particular geographic hotspots. This is essential as it allows public health officials to measure the impact of previous intervention strategies and to ensure that resources strategically target areas of concern.

## 2.2 Description of data and data source

The dataset was collected from the CDC website and the information was sourced from the National Vital Statistics System (..5…). The dataset contains rates and trends in the hypertension-related cardiovascular disease (CVD) death rates. The study period is from 2000 to 2019, with two main intervals displayed in the data 2000-2009, and 2010-2019. The data is further distributed by county, age group (35-64 years and 65 years and older), race/ethnicity, and sex.

(1= )

(2= Virani SS, et al. Heart Disease and Stroke Statistics—2020 Update: A Report From the American Heart Association. *Circulation*. 2020;141:e139–e596.)

(3= Whelton PK, et al. 2017 ACC/AHA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults. *Hypertension*. 2018;71:e13–e115)

(Mahmood SS, et al. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet*. 2014;383(9921):999–1008.)

(4= Yusuf S, et al. Modifiable risk factors, cardiovascular disease, and mortality in 155,722 individuals from 21 high-income, middle-income, and low-income countries (PURE): a prospective cohort study. *Lancet*. 2020;395(10226):795–808.)

(undefined

(5= at the following url: https://data.cdc.gov/Heart-Disease-Stroke-Prevention/Rates-and-Trends-in-Hypertension-related-Cardiovas/uc9k-vc2j/about\_data. )

## 2.3 Questions/Hypotheses to be addressed

This data analysis aimed to evaluate the rates of hypertension-related cardiovascular disease mortality in adults in the United States based on county-level data and to identify trends that might be associated with demographic and geographic factors. Furthermore, we aim to identify factors that can be used in predictive models to estimate mortality trends.

#### 2.3.0.1 Study Aims:

- Aim 1: Assess national temporal trends in hypertension-related CVD mortality, stratified by age group.

- Aim 2: Identify spatial clustering of county-level mortality.

- Aim 3: Evaluate disparities in mortality trends by race/ethnicity and sex, stratified by age group.

- Aim 4: Build predictive models using demographic and geographic features to estimate mortality trends.

#### 2.3.0.2 Hypotheses:

- Mortality trends vary significantly over time and across age groups.

- County-level mortality exhibits spatial clustering, reflecting geographic disparities.

- Temporal trends differ significantly across race and sex, with widening inequities over time.

- Predictive models can identify key demographic and spatial factors associated with higher mortality trends.

# 3. Methods

## 3.1 Data import and cleaning

We obtained the dataset from the CDC’s online portal, sourced from the National Vital Statistics System [5]. The raw CSV file was approximately 300MB in size and contained 1,103,872 rows and 24 columns. Initial preprocessing was conducted in R to remove redundant variables, reduce file size, and prepare the dataset for analysis.There were some initial processing steps performed to evaluate and delete some of the unnecessary columns and to save this data to an RDS document to reduce the size. This was all outlined in the “processingcode” document (and README file), which will not run automatically because the Excel (csv) file will not be available due to size concerns with Github uploads (available on request). This process is fully documented in the processingcode.R and processingfile-v1.qmd files within the project repository.

Non-informative or repetitive columns (e.g., columns with a single repeated value or metadata such as DataSource, Topic, Geographic\_Level) were removed. Columns describing demographic stratification were renamed for clarity: Stratification1 was renamed to age\_group, Stratification2 to race\_ethnicity, and Stratification3 to sex. The primary outcome variable, Data\_Value, was renamed mortalities.

All rows with missing values in any of the key variables (mortality rate, year, age group, race/ethnicity, sex, and geographic coordinates) were removed. The resulting cleaned dataset contained 572,572 rows and 14 variables. Summary statistics and exploratory inspection (e.g., using the skimr package) confirmed the data’s readiness for analysis. The dataset included both numeric variables (e.g., mortality rates, year, coordinates) and categorical variables (e.g., sex, race/ethnicity, age group, county).

Age groups were categorized into two broad strata: 35–64 years and 65 years and older. While this limits deeper evaluation, it facilitates comparisons across the two broad groups and supports stratified modeling approaches. The cleaned dataset was saved in .rds format for improved performance and integration with downstream analysis scripts.

All cleaning and preprocessing steps are reproducible and are logged in the repository.

## 3.2 Statistical Analysis

Statistical analysis was structured according to the study’s four primary aims and was conducted using R (version 4.3.1). All code is available in the project repository within the statistical-analysis-v2-slim.qmd script. Analyses incorporated regression modeling, spatial statistics, and supervised machine learning to assess temporal trends, demographic disparities, spatial clustering, and predictive performance.

To evaluate national temporal trends in hypertension-related cardiovascular disease (CVD) mortality from 2000 to 2019 (Aim 1), we began by fitting a linear regression model to estimate the overall change in mortality over time. We then used generalized linear models (GLMs) to assess stratified time trends by race/ethnicity, sex, and age group. Interaction terms such as Year × race/ethnicity and Year × age group were included to evaluate whether trends varied between subgroups. We estimated annual mortality slopes for each demographic category using the emtrends() function from the emmeans package and conducted pairwise comparisons using the contrast() function to formally test for differences between groups.

For spatial analysis (Aim 2), we examined county-level clustering of mortality rates using spatial statistics. County geometries were created by converting longitude and latitude coordinates into spatial features using the sf package. Spatial weights were computed using 5-nearest-neighbor relationships. Global spatial autocorrelation was assessed using Moran’s I statistic. Local clustering patterns were evaluated using Local Indicators of Spatial Association (LISA) via the spdep package. Local spatial relationships were visualized using standardized mortality z-scores and classified into high–high, low–low, or spatial outlier clusters based on the combination of mortality values and neighboring spatial lags.

To evaluate demographic disparities in temporal trends (Aim 3), we extended the GLMs used in Aim 1 by including three-way interactions between year, race/ethnicity, and sex, and stratified analyses by age group. Estimated marginal trends were obtained separately for adults aged 35–64 and those aged 65 and older. Annual slope estimates and their confidence intervals were extracted and compared across race–sex combinations. Pairwise comparisons of slopes were again conducted using emmeans, with the goal of quantifying and visualizing disparities in mortality trends across key demographic intersections.

In the final aim (Aim 4), we developed predictive models to estimate county-level mortality rates using demographic and temporal features. Two modeling approaches were implemented: a random forest model using the ranger engine and a LASSO regression model using the glmnet engine. The modeling dataset included year, age group, race/ethnicity, and sex as predictors. Categorical variables were dummy-encoded using the recipes package, and numeric predictors were normalized; zero-variance predictors were removed prior to modeling. Data were split into training (80%) and test (20%) sets, stratified by mortality outcome. Within the training set, 10-fold cross-validation was used to tune model hyperparameters. For random forests, tuning was performed over a regular grid of mtry and min\_n values. LASSO regression was tuned using a log-scaled penalty grid to identify optimal shrinkage parameters. Final model performance was evaluated on the test set using RMSE (root mean squared error), MAE (mean absolute error), and R² (coefficient of determination). Confidence intervals for these metrics were not computed due to computational constraints. Model interpretability was further assessed using variable importance plots for the random forest model and non-zero coefficient selection in LASSO.

This multi-step analytical framework enabled us to assess trends, identify disparities, detect geographic clustering, and test predictive capacity using a consistent set of demographic and temporal variables across all analytic aims.

# 4. Results

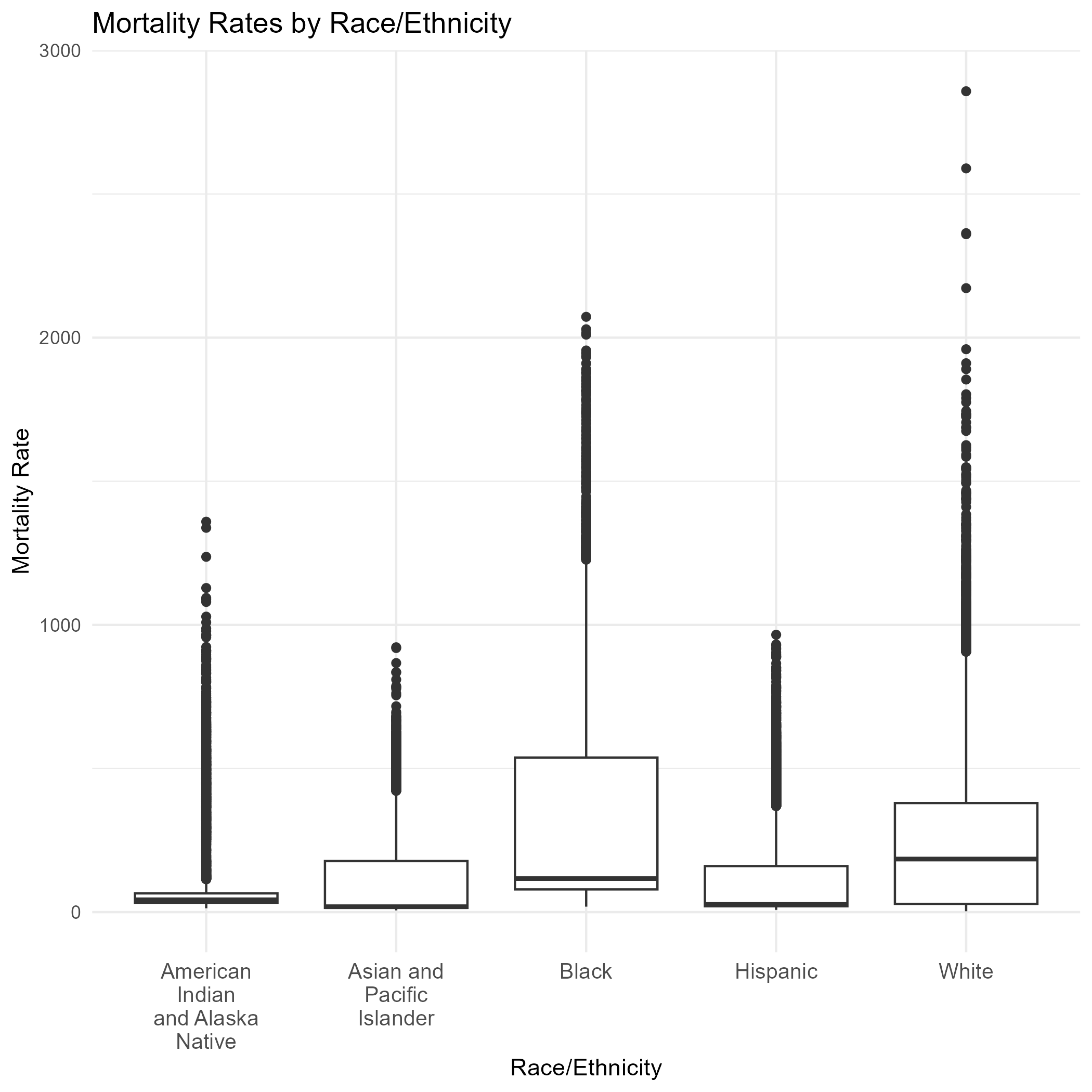
## 4.1 Descriptive and Exploratory Analysis

We began by examining overall patterns in hypertension-related cardiovascular disease (CVD) mortality among U.S. adults aged 35 and older from 2000 to 2019 (Aim 1; Hypothesis 1). Summary statistics revealed large disparities in mortality rates across racial/ethnic and age groups. For adults aged 35–64, Black individuals had the highest mean county-level mortality rate (92.2 per 100,000), followed by American Indian/Alaska Native (46.2), and White (32.9). In the 65+ age group, Black adults again had the highest mean mortality (662.6), with American Indian/Alaska Native (451.9) and White (398.3) also experiencing elevated rates.

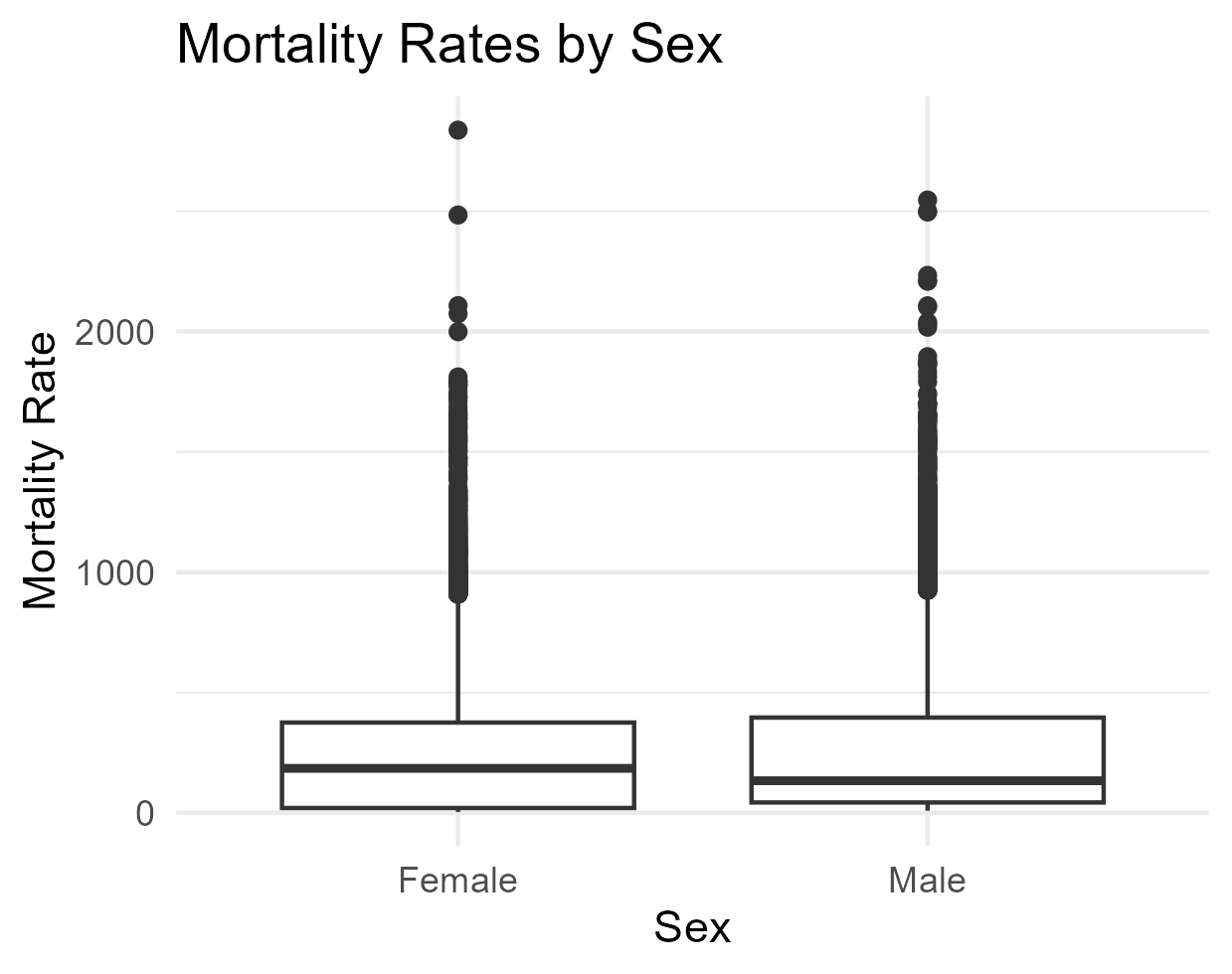
Boxplots of mortality by race/ethnicity and sex (Figures below) confirmed these disparities, showing right-skewed distributions with particularly high upper ranges for Black and American Indian/Alaska Native adults. Males had consistently higher mortality distributions than females across all groups.

Temporal visualizations indicated divergent patterns between subgroups. For example, while some populations experienced declines in mortality rates, others showed upward or stable trends. These descriptive findings supported formal modeling to assess temporal and demographic patterns.

format width height colorspace matte filesize density  
1 PNG 2100 2100 sRGB TRUE 100520 118x118



format width height colorspace matte filesize density  
1 PNG 1239 980 sRGB TRUE 37253 118x118



## 4.2 Basic statistical analysis

### 4.2.1 Temporal Trends in Mortality

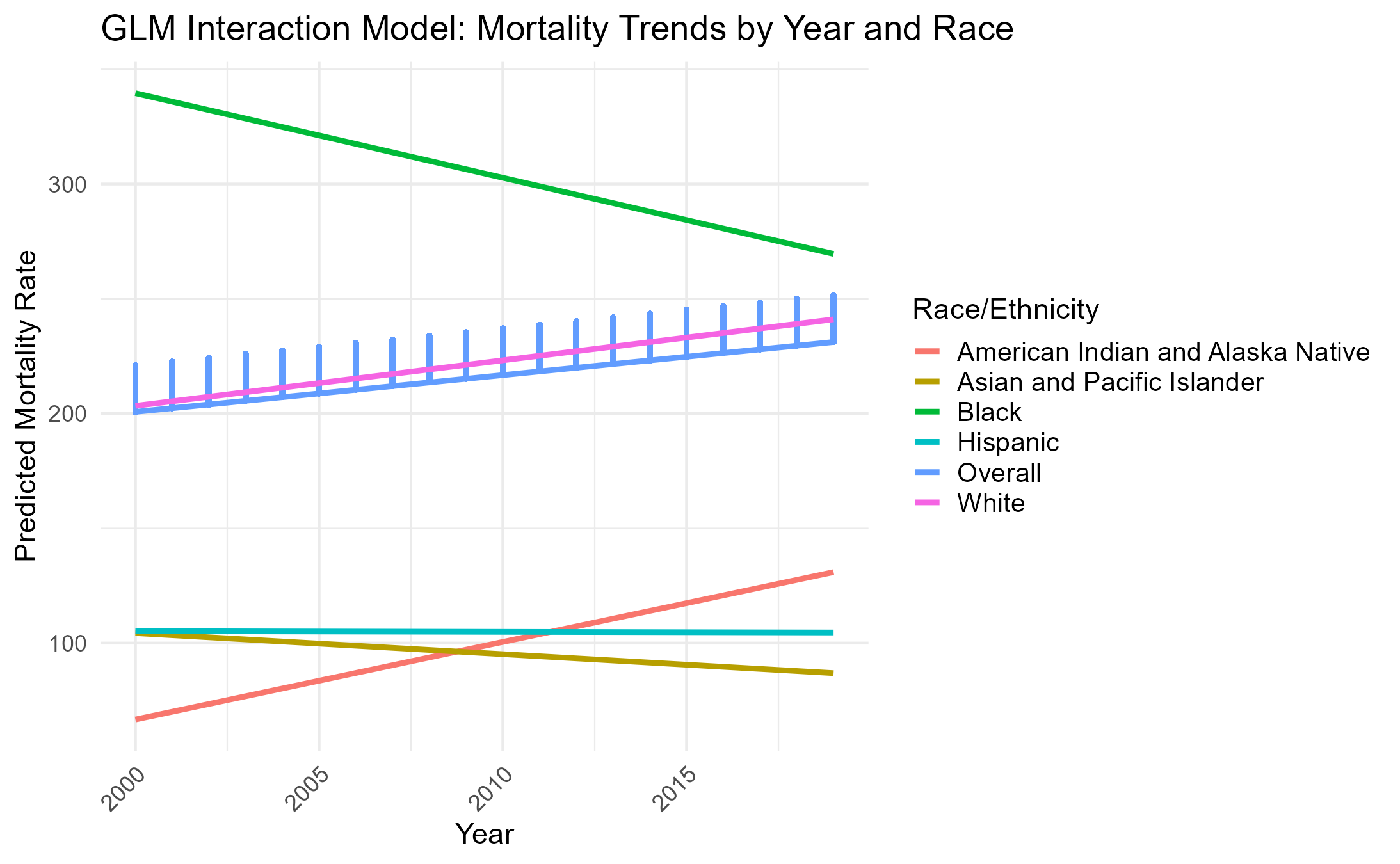
At the national level, a linear regression of overall mortality against year showed a small but statistically significant increase in mortality from 2000 to 2019 (slope = 1.40, p < 0.001). However, stratified models revealed considerable variation between groups.

A generalized linear model (GLM) incorporating race/ethnicity and sex predictors (Aim 1.2; Hypotheses 1 and 3) confirmed that Black individuals had the highest predicted mortality rates, followed by White and American Indian/Alaska Native populations. Males had significantly higher predicted rates than females across all racial/ethnic groups.

An interaction model between year and race (Aim 1.3) revealed statistically significant divergence in trends. Mortality among Black individuals declined sharply over time (slope = –7.06, p < 0.001), while American Indian/Alaska Native (slope = –3.41) and Hispanic populations showed shallower or increasing trends.

Interaction between year and age group (Aim 1.4) demonstrated that adults aged 65+ experienced a significantly steeper increase in mortality over time compared to those aged 35–64 (interaction slope = +0.22, p < 0.001). Estimated marginal trends (Aim 1.5) further confirmed this: the slope among older adults was 1.26 per year (95% CI: 1.18–1.34), versus 1.04 (95% CI: 0.96–1.12) in the younger group (p = 0.0002).

format width height colorspace matte filesize density  
1 PNG 2187 1350 sRGB TRUE 127786 118x118



### 4.2.2 Geographic Distribution and Clustering

Spatial visualizations demonstrated considerable geographic variation in mortality by race/ethnicity (Aim 2.1–2.2). Elevated mortality was most prominent among Black individuals, particularly in counties across the Southeast and Midwest. Asian and Pacific Islander populations displayed lower and more geographically diffuse mortality rates.

format width height colorspace matte filesize density  
1 PNG 4200 3000 sRGB FALSE 855811 118x118



Global spatial clustering was modest but statistically significant for the national dataset (Moran’s I = 0.0435, p < 0.001), with similar low levels of spatial autocorrelation in state-level analyses: Georgia (I = 0.0297), Mississippi (0.0289), Texas (0.0168), and Kentucky (0.0227), all p < 0.001 (Aim 2.3).

Local spatial clustering (Aim 2.4) was visualized using Local Indicators of Spatial Association (LISA). Statistically significant high–high clusters were identified in counties across parts of the Southeast, but the majority of counties showed no local spatial clustering. This suggests that while some mortality hotspots exist, broad demographic and structural drivers likely contribute more to the observed geographic distribution than tightly bounded spatial clusters.

### 4.2.3 Demographic Disparities in Temporal Trends

We assessed demographic disparities in mortality trends across race/ethnicity and sex within age groups (Aim 3.1–3.3; Hypothesis 3). Among adults aged 35–64, Black individuals experienced the steepest annual decline in mortality (–3.68), while American Indian/Alaska Native individuals showed the largest increases (+3.38). Overall, males experienced steeper increases than females.

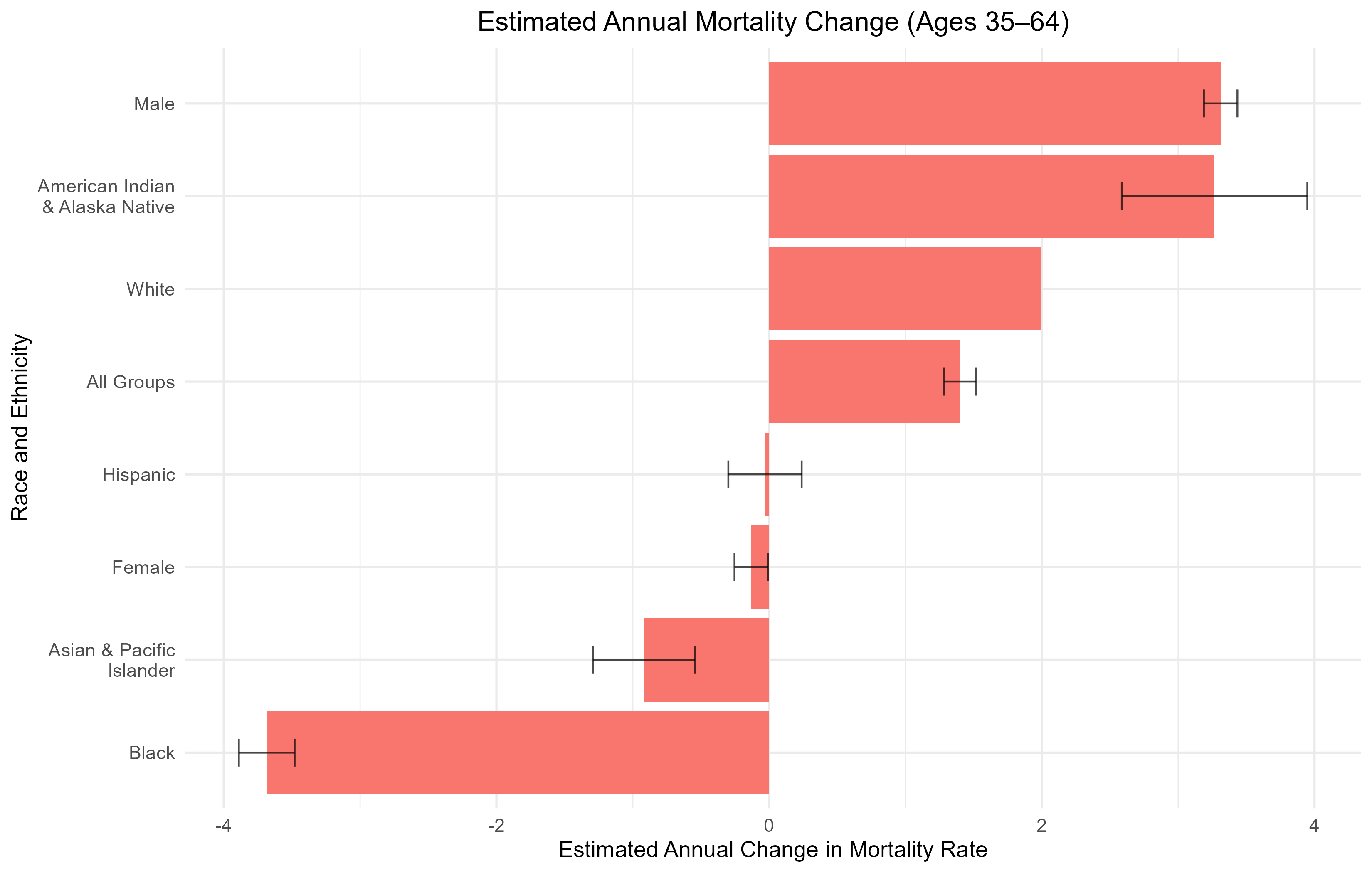
Pairwise comparisons showed the greatest differences between Overall Male and Black Overall groups (+7.00 per year), followed by contrasts involving American Indian/Alaska Native, White, and Asian/Pacific Islander groups. These results indicate persistent and widening disparities in mortality trends, particularly among marginalized populations.

Age Group Estimate SE CI  
1 Ages 35-64 years 1.036 0.041 0.95–1.12  
2 Ages 65+ years 1.260 0.043 1.17–1.34

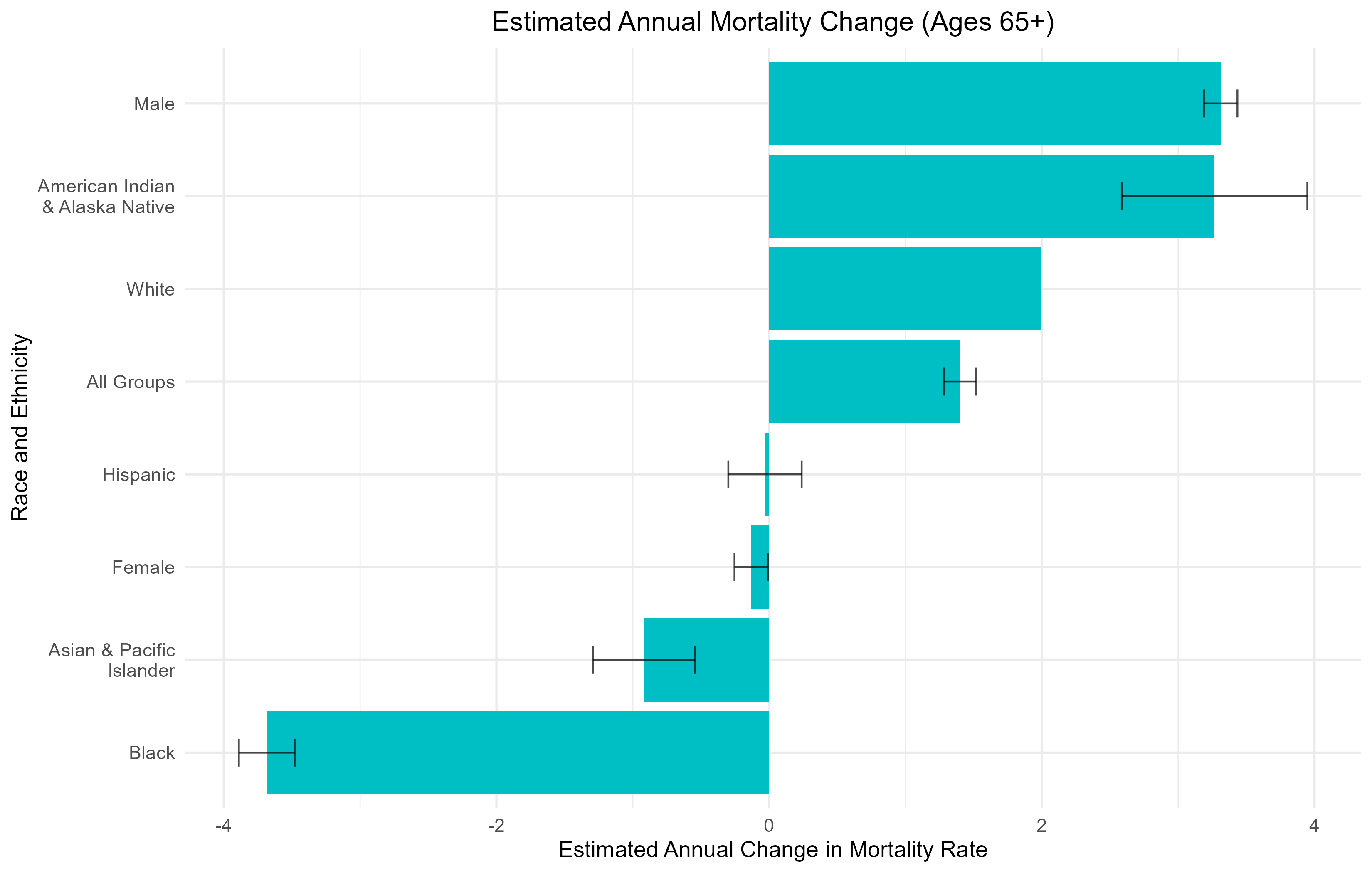
Race Sex Estimate SE CI  
1 American Indian and Alaska Native Female 3.380 0.471 2.46–4.3  
2 Asian and Pacific Islander Female -0.918 0.361 -1.63–-0.21  
3 Black Female -3.685 0.197 -4.07–-3.3  
4 Hispanic Female -0.030 0.259 -0.54–0.48  
5 Overall Female 1.601 0.084 1.44–1.77  
6 White Female 1.983 0.116 1.76–2.21  
7 American Indian and Alaska Native Male 3.380 0.471 2.46–4.3  
8 Asian and Pacific Islander Male -0.918 0.361 -1.63–-0.21  
9 Black Male -3.685 0.197 -4.07–-3.3  
10 Hispanic Male -0.030 0.259 -0.54–0.48  
11 Overall Male 1.601 0.084 1.44–1.77  
12 White Male 1.983 0.116 1.76–2.21  
13 American Indian and Alaska Native Overall 3.380 0.471 2.46–4.3  
14 Asian and Pacific Islander Overall -0.918 0.361 -1.63–-0.21  
15 Black Overall -3.685 0.197 -4.07–-3.3  
16 Hispanic Overall -0.030 0.259 -0.54–0.48  
17 Overall Overall 1.601 0.084 1.44–1.77  
18 White Overall 1.983 0.116 1.76–2.21

Additional visualizations from Aim 3.4 reinforce these findings:

format width height colorspace matte filesize density  
1 PNG 3300 2100 sRGB TRUE 118915 118x118



format width height colorspace matte filesize density  
1 PNG 3300 2100 sRGB TRUE 117383 118x118



### 4.2.4 Predictive Modeling of Mortality Rates

To evaluate the ability of demographic and temporal factors to predict hypertension-related CVD mortality at the county level, we developed and compared two supervised machine learning models: a Random Forest and a LASSO regression model (Aim 4; Hypothesis 4). Both models were trained on 80% of the dataset and evaluated on a held-out 20% test set, with 10-fold cross-validation used during training to optimize hyperparameters and reduce overfitting.

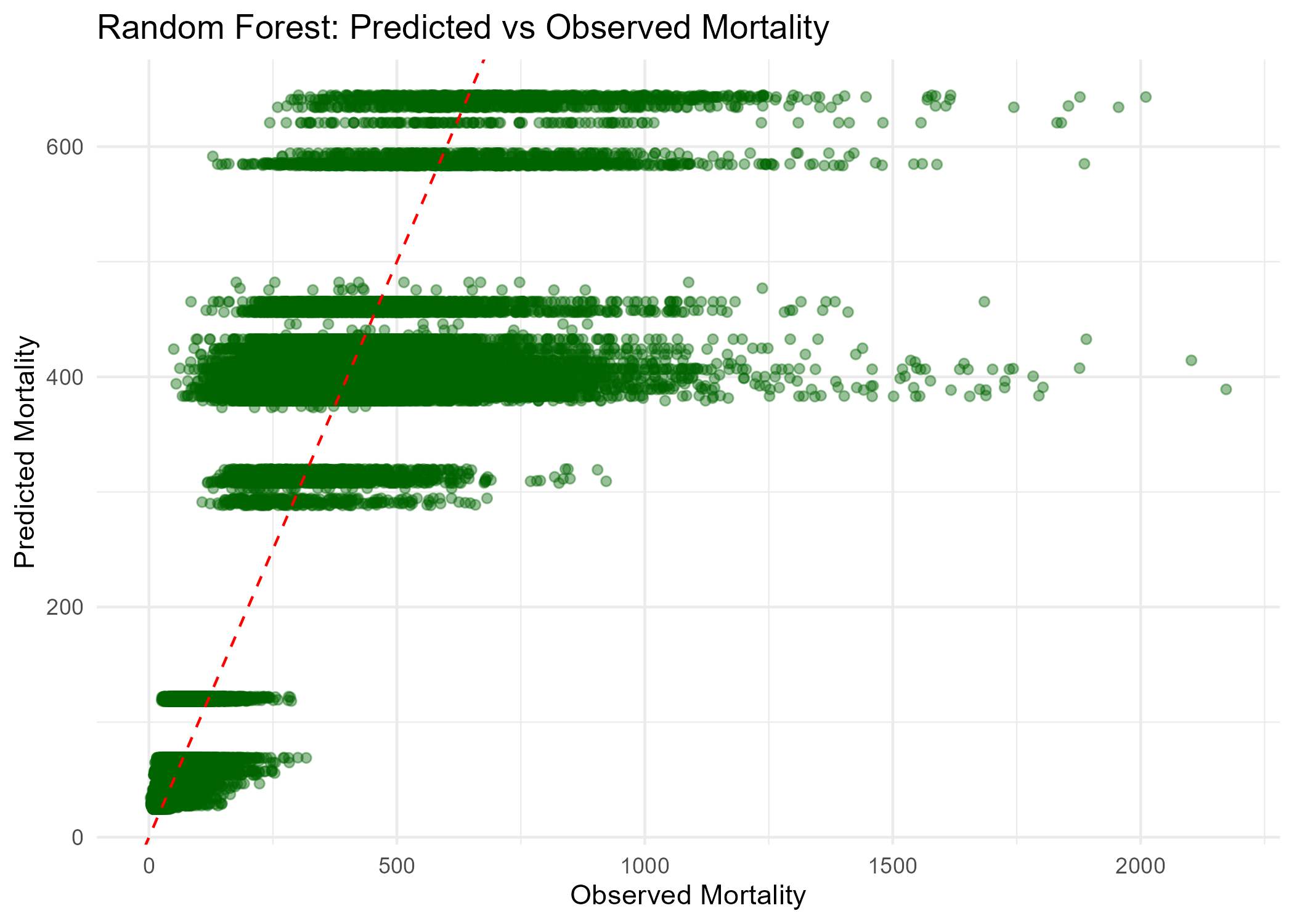
The Random Forest model demonstrated strong predictive accuracy. The best-performing model used all four available predictors (year, age group, race/ethnicity, and sex) with a minimum node size of 15. On the test set, this model achieved a root mean squared error (RMSE) of 115.3, a mean absolute error (MAE) of 66.9, and an R² of 0.748, indicating that approximately 75% of the variance in county-level mortality was explained by the model. Variable importance scores revealed that race/ethnicity and year were the most influential predictors, followed by age group and sex. The predicted versus observed plot showed generally good alignment, though some dispersion at the extremes suggested modest underfitting in counties with very high or low mortality.

Table 1: Model Performance Summary

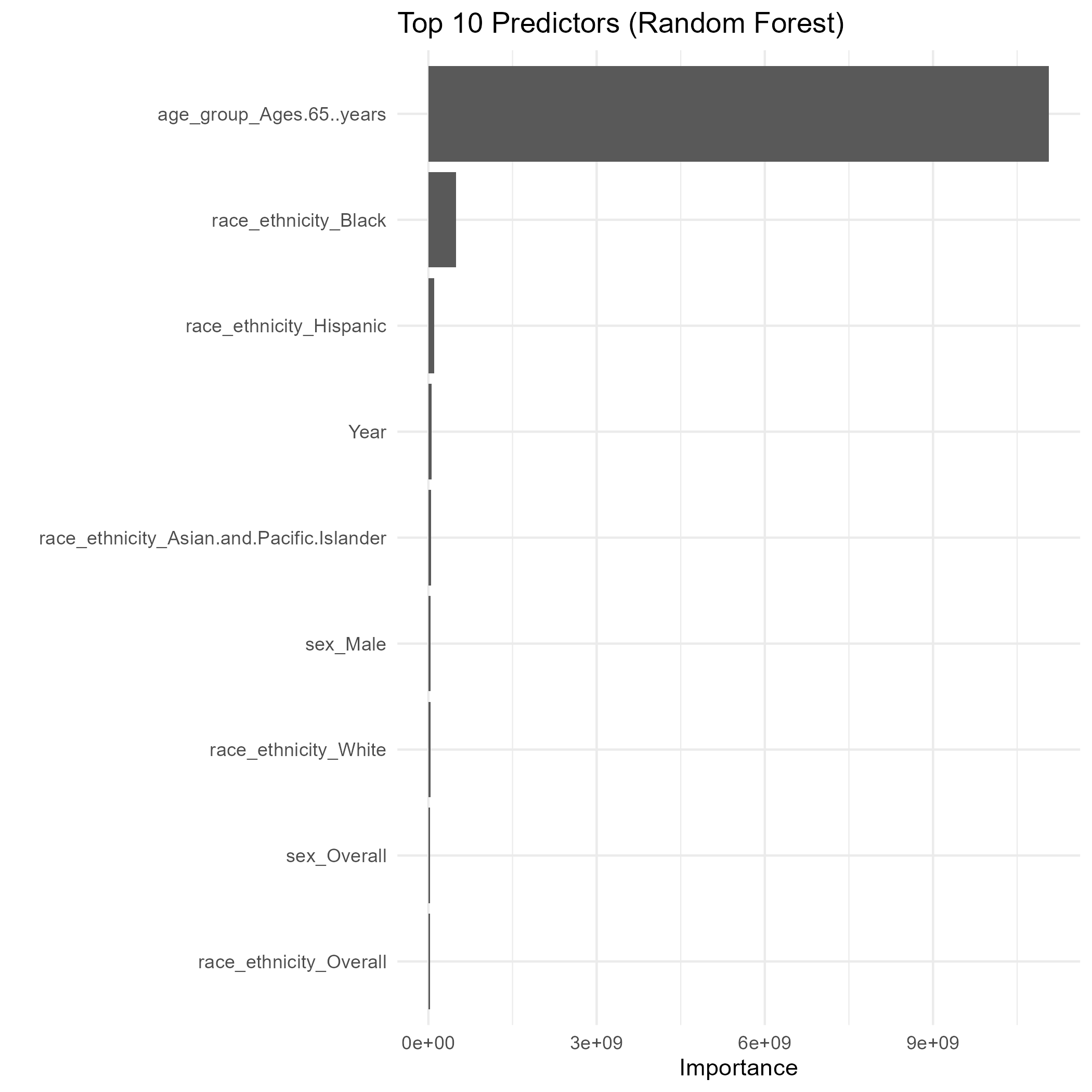
Test Set Evaluation for Random Forest and LASSO

| Model | RMSE | R² | MAE |
| --- | --- | --- | --- |
| Random Forest | 115.3 | 0.75 | 66.86 |
| LASSO Regression | 119.54 | 0.73 | 70.66 |

# A tibble: 1 × 7  
 format width height colorspace matte filesize density  
 <chr> <int> <int> <chr> <lgl> <int> <chr>   
1 PNG 2100 1500 sRGB TRUE 442545 118x118

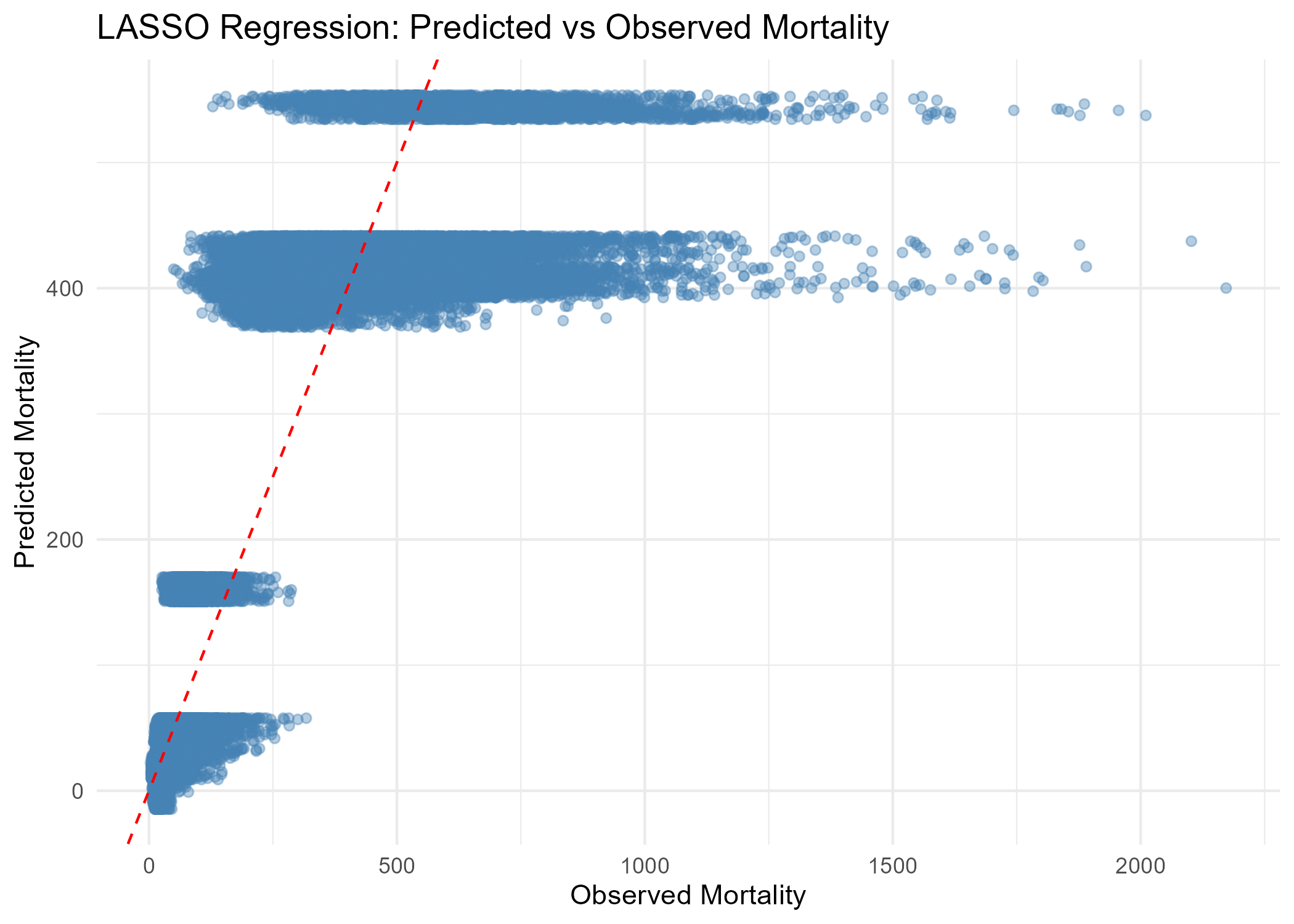


# A tibble: 1 × 7  
 format width height colorspace matte filesize density  
 <chr> <int> <int> <chr> <lgl> <int> <chr>   
1 PNG 2100 2100 sRGB TRUE 93952 118x118



The LASSO regression model provided a more interpretable, though slightly less accurate, alternative. On the same test set, the LASSO model yielded an RMSE of 119.5, an MAE of 70.7, and an R² of 0.726. This model performed coefficient shrinkage and variable selection, shrinking several dummy-coded predictors toward zero. While this feature enhances interpretability and highlights influential variables, it also limits the model’s flexibility to capture complex non-linear patterns. The predicted versus observed plot for the LASSO model revealed horizontal banding patterns. This reflects the impact of dummy-encoded categorical variables — many counties shared identical predictor combinations, resulting in repeated predictions and horizontal clustering in plots. These effects were more pronounced in LASSO than in the Random Forest, which is better able to handle categorical interactions.

# A tibble: 1 × 7  
 format width height colorspace matte filesize density  
 <chr> <int> <int> <chr> <lgl> <int> <chr>   
1 PNG 2100 1500 sRGB TRUE 470333 118x118



Despite these differences, both models demonstrated solid performance in capturing patterns in the data. The Random Forest provided better predictive accuracy, while the LASSO model offered greater clarity in identifying relevant features. Taken together, these results indicate that demographic and temporal factors are strong predictors of mortality variation at the county level, supporting Hypothesis 4 and offering tools for future public health risk modeling.

\_

# 5. Discussion

## 5.1 Summary and Interpretation

We examined patterns and predictors of hypertension-related cardiovascular disease (CVD) mortality among U.S. adults aged 35 and older between 2000 and 2019. Through a combination of exploratory visualization, statistical modeling, spatial clustering, and machine learning, we found consistent evidence of disparities in mortality across time, geography, and demographic subgroups. These findings underscore the ongoing burden of hypertension-related CVD in the United States and highlight areas for targeted intervention.

National trends revealed that overall mortality rates remained relatively stable over the study period. However, stratified models showed considerable variation between groups. Black individuals experienced the highest mortality rates throughout, while American Indian/Alaska Native populations showed a concerning upward trend in recent years. Adults aged 65 and older had significantly steeper increases in mortality rates than younger adults. These findings are consistent with previous national surveillance reports that have identified persistent disparities in CVD burden across racial/ethnic and age groups [2,4].

Spatial analyses detected statistically significant but modest spatial clustering of mortality at the national and state levels. While global Moran’s I statistics confirmed the presence of spatial autocorrelation, the magnitude was small (I = 0.0435 nationally). Local spatial clustering (LISA) identified limited high-mortality hotspots, primarily concentrated in the Southeast. These results suggest that broader structural and demographic factors—rather than localized geographic effects—may be the primary drivers of hypertension-related CVD mortality. This is in line with previous studies indicating that health disparities are often influenced more by social determinants and systemic inequities than by geographic proximity alone [4,6].

Demographic subgroup analyses revealed widening inequities in mortality trends. Among adults aged 35–64, Black individuals experienced the steepest annual declines, suggesting some gains in this group. However, these improvements occurred in parallel with rising mortality among American Indian/Alaska Native individuals and males, who showed the largest annual increases. Among adults aged 65 and older, mortality increased for nearly all groups. These divergent trends highlight the importance of considering intersecting axes of inequality—such as race, sex, and age—in public health surveillance. They also support prior research showing that demographic-specific interventions are necessary to effectively reduce disparities in chronic disease outcomes [3,4,7].

In the predictive modeling component, we assessed whether demographic and temporal variables could accurately predict county-level mortality. The Random Forest model performed best, explaining approximately 75% of the variation in mortality rates (R² = 0.748), with an RMSE of 115.3. Race/ethnicity and year were the most influential predictors. The LASSO regression model, while slightly less accurate (R² = 0.726), offered greater interpretability by selecting a reduced subset of predictors and shrinking non-influential coefficients to zero. These results suggest that relatively simple demographic and temporal features can provide strong predictive power, and that different modeling approaches offer complementary strengths. The Random Forest’s flexibility allowed for more accurate predictions, while LASSO highlighted core factors driving variability.

Notably, both models showed horizontal banding in predicted vs. observed plots—an artifact of dummy-encoded categorical variables and repeated predictor combinations across counties. This highlights a broader methodological challenge in population-level modeling using grouped categorical features. Although machine learning approaches can improve predictive performance, careful attention must be paid to encoding strategies and interpretability.

These findings contribute to a growing body of evidence emphasizing the importance of demographic stratification and predictive modeling in public health research. In particular, our analysis supports previous work demonstrating the utility of ensemble and regularized regression models in epidemiologic prediction tasks [8,9]. Importantly, our results reinforce that even well-known risk factors—such as age, sex, race/ethnicity, and time—continue to explain a large proportion of the variance in hypertension-related CVD mortality and may be sufficient for initial risk stratification at the population level.

## 5.2 Strengths and Limitations

This study benefits from the use of a comprehensive, nationwide dataset covering nearly two decades of mortality records. The integration of statistical and machine learning methods allowed us to not only describe mortality patterns but also assess their predictability. However, several limitations should be noted. First, the analysis was limited to county-level aggregated data, which may obscure within-county heterogeneity and limit causal inference. Second, although we used age-standardized mortality rates, residual confounding by comorbidities, access to care, socioeconomic status, or other contextual factors cannot be excluded. Lastly, while our models achieved high predictive accuracy, they were not designed to support individual-level predictions or establish causal relationships.

## 5.3 Implications and Future Directions

Our findings highlight important and persistent disparities in hypertension-related CVD mortality that should inform ongoing public health surveillance and intervention efforts. Geographic clustering, although weak, may still guide resource allocation to high-burden areas. Predictive models developed here can support proactive health planning, particularly when integrated with local surveillance systems. Future research should explore integrating additional variables—such as healthcare access, environmental exposures, and social determinants of health—to further refine predictions and address upstream drivers of CVD disparities.

## 5.4 Conclusions

This study provides a comprehensive analysis of hypertension-related cardiovascular disease (CVD) mortality among U.S. adults aged 35 and older from 2000 to 2019, using national county-level data. Our findings reveal that while overall mortality trends remained relatively stable at the national level, significant disparities persist across demographic and geographic groups. Black and American Indian/Alaska Native populations consistently experienced higher mortality rates, with the steepest increases observed among older adults and males. Spatial analysis identified weak but statistically significant clustering, particularly in the Southeastern United States, though most variation was driven by broader structural and demographic patterns rather than tightly localized effects.

Temporal modeling demonstrated that subgroup-specific trends diverged substantially, suggesting that recent gains in cardiovascular health have not been equitably distributed. Our predictive modeling further supports this, with demographic features—particularly race/ethnicity and age—emerging as dominant predictors of county-level mortality risk. Random Forest models outperformed LASSO regression in accuracy, but both approaches reinforced the critical role of social and demographic context in shaping health outcomes.

Taken together, these findings underscore the urgent need for targeted public health strategies that account for longstanding inequities in cardiovascular risk. Future research should integrate social determinants of health, healthcare access, and policy-level factors to better understand the drivers of these disparities and to design interventions that are both equitable and effective.

Note that this cited reference will show up at the end of the document, the reference formatting is determined by the CSL file specified in the YAML header. Many more style files for almost any journal [are available](https://www.zotero.org/styles). You also specify the location of your bibtex reference file in the YAML. You can call your reference file anything you like.

# 6. References