Exploring the Health Inequalities: The Interplay of Climate, Race, and Education on Epigenetic Aging in HIV-Affected Populations Using Deep Learning

Cong Cao [1,2], Ryan Hu[1], Elizabeth C. Breen[3], Roger Shih[4], Mary E. Sehl[9], Frank Palella[3], Matthew Mimiaga[4], Jeremy Martinson[5], Todd Brown[6], Sheri D. Weiser, B[7], Elizabeth Crabb Breen [8], R. Michael Alvarez[1], Beth D. Jamieson[8], Christina M. Ramirez[9]

January 20, 2025

Affiliations:

[1] California Institute of Technology, USA

[2] Norwegian University of Science and Technology, Norway

[3] Potocsnak HIV and Aging Center within the Potocsnak Longevity Institute, Northwestern Memorial Hospital, NMH/Feinberg, 676 N. St Clair Suite 940, Chicago, Illinois 60611

[4] UCLA Fielding School of Public Health and Psychiatry and Biobehavioral Sciences at UCLA David Geffen School of Medicine, Los Angeles, CA

[5] Department of Infectious Diseases and Microbiology, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, USA

[6] Fashion Institute of Technology

[7] UC San Francisco's Division of HIV, Infectious Diseases and Global Medicine at Zuckerberg San Francisco General Hospital

[8] Division of Hematology-Oncology, Department of Medicine, David Geffen School of Medicine at UCLA, University of California Los Angeles, Los Angeles, California 90095

[9] Department of Biostatistics, Fielding School of Public Health, University of California Los Angeles, Los Angeles, California 90095, USA

[10] Department of Microbiology & Immunology in the UCLA School of Medicine, Los Angeles, California 90095, USA

Corresponding author: Cong Cao; congc@caltech.edu

Abstract

This study investigates the relationship between climate change, air pollution, health inequality, and methylation modifications that affect gene expression, with a particular focus on whether these factors have an interactive impact on aging in people living with HIV (PLWH) more than in those without HIV. We used deep learning techniques on data from four U.S. cities, including Los Angeles, to better understand the impact of climate change on individuals living with HIV, with the goal of informing climate and public health policies. To explore the relationships between climate variables and regional characteristics, we introduced interaction terms into our models, which revealed notable disparities in health outcomes across the cities. Our regression analysis showed strong correlations between climate variables and health outcomes in Chicago and Los Angeles, whereas the associations were significantly weaker in Pittsburgh. Specifically, climate factors like precipitation and temperature had minimal effects on epigenetic aging. However, regional characteristics such as racial composition were found to negatively correlate with aging levels, with higher percentages of white residents associated with slower epigenetic aging. Additionally, education was positively linked to better outcomes, particularly in Pittsburgh. The model suggests other unmeasured factors may be influential, despite the limited role of climate variables. Additionally, there were minimal differences in predictive performance between deep learning and linear regression, suggesting that traditional regression can effectively capture data patterns. In conclusion, this research highlights the significant roles of regional variables, racial composition, and education on epigenetic aging while underscoring the minimal influence of climate factors. These findings emphasize the need to address health inequalities and develop nuanced policies considering the complex interplay of environmental and social factors in health outcomes.

Keywords

Deep learning, aging, health inequities, prediction, climate

1. Introduction

People living with HIV (PLWH) experience accelerated epigenetic aging compared to age-matched individuals without HIV (PLWoH) [1]. Various studies have investigated how DNA methylation age and environmental factors play a role in environmental epidemiology [2]. Environmental exposures, such as meteorological conditions, air pollutants, and socioeconomic factors, all play a significant role in health outcomes. Socioeconomic factors, including BMI, education, and smoking, are closely linked to health, with individuals from lower socioeconomic statuses often experiencing poorer health outcomes and shorter life expectancy. Socioeconomic factors could lead to health inequalities, which reflect the systematic differences in health and access to resources among different social groups, evident in disease rates, health behaviors, and healthcare usage. Importantly, epigenetic and environmental factors do not operate in isolation. The individual impacts of air pollution, weather conditions, health inequalities, and genetic factors, along with their interactions, collectively shape the health effects of epigenetic aging.

Climate change is causing extreme weather, like higher temperatures, which can create more health risks for older people. Other changes from climate change, like lower humidity, changing rainfall, and increased wind speeds, are also important when studying aging in people with HIV. Air pollution, especially fine particles, and ground-level ozone, harms the health of HIV/AIDS patients, showing the need for better health policies and regulations [31-34].

Previous research has mainly looked at how individual environmental factors, like climate and air pollution, affect health outcomes separately. However, the combined effects of these factors on aging in people living with HIV (PLWH) are not well understood. Studies have shown that long-term exposure to pollutants like PM2.5 can be linked to aging-related diseases, but its effect on epigenetic aging is unclear [3]. This is especially important for PLWH, as they may age faster due to both the virus and environmental stressors. In this study, we aim to explore how environmental factors, like temperature, precipitation, and air pollution, together with socio-demographic factors, affect epigenetic aging in PLWH. We hypothesize that HIV, combined with environmental exposures, leads to faster epigenetic aging and that factors like racial composition and education level can influence this process. Our study builds on previous research, such as Olatosi et al. (2021), by considering multiple factors and their interactions.

Our study will analyze existing data on aging rates in men living with and without HIV, utilizing data from the MACS/WIHS Combined Cohort Study (MWCCS). We will integrate this cohort data with historical air pollution and weather data from the regions where participants live, to better understand the relationship between environmental factors and aging rates. Four cities in the U.S. will be analyzed, including Los Angeles, to develop relevant policy suggestions. Previous research has looked at specific areas, like Fulton County in Georgia, to understand the factors leading to the AIDS epidemic [4]. However, different cities may have communities with varying characteristics, and we will explore whether these differences significantly impact people living with HIV. Orel et al. (2022) identified key factors for predicting HIV status in Africa, such as location, age, number of sexual partners, condom use, and wealth index. However, this study has limitations,

including differences between countries or locations [5].

Current research on genetic variation and economics faces challenges due to small sample sizes and excessive variables, often leading to overfitting and increased computational complexity in traditional statistical methods. In contrast, this study will employ deep learning techniques to predict this relationship and explore health inequality. Deep learning methods are innovative because they use artificial intelligence to develop algorithms and statistical models that can learn from data and make predictions or decisions based on that information [6], [7]. It is important for several reasons: 1. Data-driven insights: Enables us to extract valuable insights and patterns from large and complex datasets—such as the MWCCS—that would be challenging or impossible to uncover through traditional statistical methods; 2. Automation and efficiency: It can automate tasks and processes that would otherwise require significant human effort and time, leading to increased efficiency, productivity, and cost savings; and 3. Predictive capabilities: Makes accurate predictions and forecasts about future events or outcomes. Through statistical analysis and deep learning algorithms, we will explore the role of these factors in explaining the results of different measures of epigenetic aging. These analyses will provide a deep understanding of how an individual's socio-demographics affect their healthy epigenetic aging process and help develop more effective health intervention strategies.

Our study will explore the impact of weather and air pollution variables, other than heat waves, on epigenetic aging, with a focus on how these factors interact to influence HIV aging, using deep learning methods to analyze large datasets. We will examine four U.S. cities. The goal is to provide insights that inform climate and public health policies aimed at mitigating the effects of climate change on aging in PLWH.

Related work

Climate change is causing extreme heat, which may pose greater health risks for the elderly. Other weather factors are also crucial for studying epigenetic aging in people living with HIV (PLWH). Air pollution, especially fine particulate matter (PM2.5 and PM10) and ground-level ozone (O3), has been shown to negatively impact the mortality of HIV/AIDS patients, highlighting the need for health policy changes [31-34]. While some studies have looked at the effects of air pollution, genetics, smoking, and BMI on those living with HIV/AIDS, none have used machine learning to analyze these interactions [16, 17]. Previous research has explored socio-behavioral factors affecting HIV prevalence in sub-Saharan Africa, however, no specific factors have been identified for men in isolation; studies focusing solely on men are limited. [16, 17]. We will investigate whether different U.S. cities, with their unique community characteristics, significantly influence individuals living with HIV/AIDS.

Extreme weather events, such as rising temperatures, are becoming more frequent and are expected to pose greater health risks to the elderly compared to the general population. Many studies have confirmed that heat waves can lead to increased mortality and have differential health consequences. For example, a 2020 report by *The Lancet* proposed that, as a group, the elderly are vulnerable to climate change due to poor adaptability to extremely high temperatures [8]. However, other meteorological factors caused by climate change are also important for research on epigenetic aging in PLWH. In addition to heat, global warming will lead to reduced

relative humidity [9], increased air pressure, changes in precipitation [10], increased global wind speeds [11], and alterations in shortwave radiation [12]. Air pollution, particularly PM2.5, PM10, and O3, has been shown to negatively affect the mortality of HIV/AIDS patients, further highlighting the urgent need for changes in health policies and regulations.

The adverse effects of air pollution, particularly fine particulate matter (PM2.5 and PM10) and ground-level ozone (O3), on HIV/AIDS patient mortality have also been documented, emphasizing the need for health policy changes and regulations [13]. Liang et al. (2024) explored the correlation between prognosis and air pollution (specifically PM2.5 and PM10) among HIV/AIDS patients in Wuhan, integrating satellite data and ground-based measurements. Employing space-time extremely randomized trees (STET) models, they analyzed the relationship between PM concentrations and AIDS-related deaths and complications using time-varying Cox proportional hazard models. Their findings revealed that long-term exposure to PM1, PM2.5, and PM10 was positively associated with increased risks of AIDS-related complications and mortalities [13]. Some studies have explored the impact of air pollution, epigenetics, smoking, and BMI on individuals living with HIV/AIDS, yet none have utilized machine-learning techniques to analyze these interactive relationships. Zhang et al. (2023) analyzed the long-term effects of particulate matter (PM2.5 and PM10) on HIV/AIDS patients in Hubei province, China, from 2010 to 2019, utilizing Cox proportional hazards models. They found that each 1 μg/m³ increase in PM2.5 and PM10 was associated with a 1.65% and 0.90% increased risk of AIDS-related deaths, respectively, with stronger associations in older patients. This research highlights the potentials for applying machine learning techniques, as such methods were not employed [14]. Chen et al. (2024) investigated the impact of ground-level ozone (O3) on people living with HIV undergoing antiretroviral therapy in Guangxi, China, from 2012 to 2019, using a longitudinal study with a generalized linear mixed effects model. They reported that a 10 μg/m³ increase in ambient O3 concentration was associated with increased mortality odds, underscoring the necessities for preventive measures to reduce O3 exposure. The absence of machine learning techniques in this study also indicates a valuable opportunity for future research [15].

Different cities likely have distinct communities with varying characteristics, and we will explore whether these differences significantly affect individuals living with HIV/AIDS. Some research has examined specific locales, such as Fulton County, Georgia, and diverse populations to understand the factors contributing to the HIV epidemic [4]. Certain communities experience elevated rates of HIV/AIDS incidence and complications, potentially due to factors such as smoking and obesity, underscoring the necessity for targeted interventions [18, 19]. Ordonez and Marconi (2012) highlighted the contributions of ethnicity, poverty, gender relations, and geographic regions to the epidemic among vulnerable populations. Despite a global decline in HIV rates, certain communities still face rising incidences, indicating a need for intervention. A comparative analysis across cities would be beneficial, as Los Angeles may have different racial demographics and poverty rates compared to cities like Atlanta, which could unveil important relationships [20]. Previous studies have employed machine learning models to predict HIV risk and care outcomes using Bayesian Networks, Random

Forests, and Neural Networks [3]. The varied models have produced differing conclusions, with Bayesian Network models effectively predicting HIV care status based on variables such as years in care and year of diagnosis, while gradient-boosting tree models excelled in predicting HIV status based on factors like geographic location and sexual behavior. Despite these advancements, limitations remain, as many studies necessitate further validation across different regions and overlook temporal relationships among variables [21]. Orel et al. (2022) investigated high-yield HIV testing methods by predicting HIV status based on socio-behavioral characteristics in East and Southern Africa. They tested four machine learning algorithms, identifying key predictive variables such as geographic location, age, number of sexual partners, condom usage, and wealth index. However, the study's limitations include variations between countries and reliance on survey data [5]. Skaathun et al. (2021) examined the relationship between geography and HIV transmission in Los Angeles County (LAC) by analyzing de-identified surveillance data from 8,186 HIV-positive individuals between 2010 and 2016. They constructed a genetic transmission network using HIV-TRACE, allowing them to evaluate geographic assortativity, time-space concordance, and the Jaccard coefficients. Their findings indicated that genetic clustering serves as a more effective indicator of HIV transmission patterns than time-space clustering. We will involve geographical data analyses, particularly in Los Angeles County [22].

Regarding epigenetic aging, we will evaluate five measures of epigenetic age using weighted elastic net regression models based on methylation levels at specific CpGs from our Infinium Methylation EPIC BeadChip method. The measures include Horvath's Pan-tissue age (353 CpGs), Extrinsic age (71 CpGs) [23], Phenotypic Age (513 CpGs) [24], Grim Age (1,030 CpGs) [25], and Skin & Blood Age (391 CpGs) [26]. These measures are commonly used epigenetic clocks that estimate biological age based on DNA methylation patterns, reflecting aging processes and health risks. Horvath's Pan-tissue Age (353 CpGs) is a universal clock applicable to multiple tissue types; Extrinsic Age (71 CpGs) focuses on immune system-related aging; Phenotypic Age (513 CpGs) combines clinical biomarkers to assess disease risk and mortality; Grim Age (1,030 CpGs) predicts lifespan and health outcomes; and Skin & Blood Age (391 CpGs) is optimized for skin and blood tissues. For each epigenetic clock, we will calculate a rate of aging by determining the ratio of the change in DNA methylation (DNAm) age between Visit 2 and Visit 1 to the change in chronological age between these visits [27]. Additionally, we will compute a shortening rate for DNAmTL by evaluating the ratio of the change in DNAmTL between Visit 2 and Visit 1 to the difference in chronological age for the same visits. To assess the effective sample size, our data will replicate that used in Sehl et al. (2022) [28, 29]. Xing and Poslad (2021) investigated the relationship between geography and HIV transmission clusters in Los Angeles County, revealing that while molecular clustering provides valuable insights into HIV transmission dynamics, its association with geographic patterns is relatively weak in urban environments. This suggests that genetic clustering may serve as a better indicator of transmission patterns compared to traditional time-space clustering methods in such contexts [30].

2 Methods

In this paper, we will correlate existing genome-wide epigenetic data with air pollution and meteorological data from publicly available sources, which have high temporal frequency and high spatial resolution. The epigenetic data is derived from men enrolled in the Multicenter AIDS Cohort Study (MACS). We have eight hundred samples. The earliest samples date back to 1984, and the latest samples were collected in 2019. Compared with other types of data, the advantage of MWCCS-based data is that it contains well-characterized comprehensive biologic, socioeconomic longitudinal data. The variation and volume of this dataset are ideal for machine learning. Another advantage of MWCCS-based data is that it is geographically informed and therefore can be easily linked to environmental data. We use data from the Multicenter AIDS Cohort Study (MACS) collected since 1984. Currently, 0.745% of our participants identify as white, while 0.255% identify as non-white. Nevertheless, the current study will incorporate other socioeconomic and demographic information available in the data. Air quality and climate data were collected from the United States Environmental Protection Agency (EPA), the National Centers for Environmental Information (NCEI), and the Atmospheric Composition Analysis Group at Washington University in St. Louis.

After normalizing the data, we conducted feature engineering. We generated interaction terms by multiplying individual variables to see how they influence the outcome. This process resulted in a dataset containing 2,092 variables and 800 observations.

2.1 Data description

Tables 1 and 2 present a statistical summary of various environmental, health, and demographic variables, offering insights into their distributions and central tendencies. The environmental variables considered include precipitation, temperature, carbon monoxide (CO), nitrogen dioxide (NO2), ozone, and sulfur dioxide (SO2). Together with Figure 1, we find that SPWPM2.5 and SBMPM2.5 exhibit minimum values of 0.0 and 0.0 μg/m3, respectively, with maximums of 16.6 and 17.2 μg/m3. The means slightly exceed the medians, suggesting a right skew in their distributions, potentially reflecting episodes of poor air quality that lead to spikes in PM2.5 concentrations. It is important to note that SPWPM2.5 and SBMPM2.5 contain a significant number of missing values, indicating variations in PM2.5 concentrations from different sources.

Carbon monoxide (CO) has a minimum of 0 and a maximum of 5.5 ppm, with an average of 0.90 ppm, indicating generally low levels but acute increases during specific activities like traffic or industrial processes. Nitrogen dioxide (NO2) concentrations range from 0.0 to 149.2 ppb, with a mean of 39.3 ppb, highlighting exposure risks in urban areas with higher vehicle emissions. Ozone levels vary from 0.034 to 0.115 ppm, while sulfur dioxide (SO2) levels range from 0.0 to 99.0 ppb, underscoring potential health risks associated with ozone exposure, particularly in warmer months.

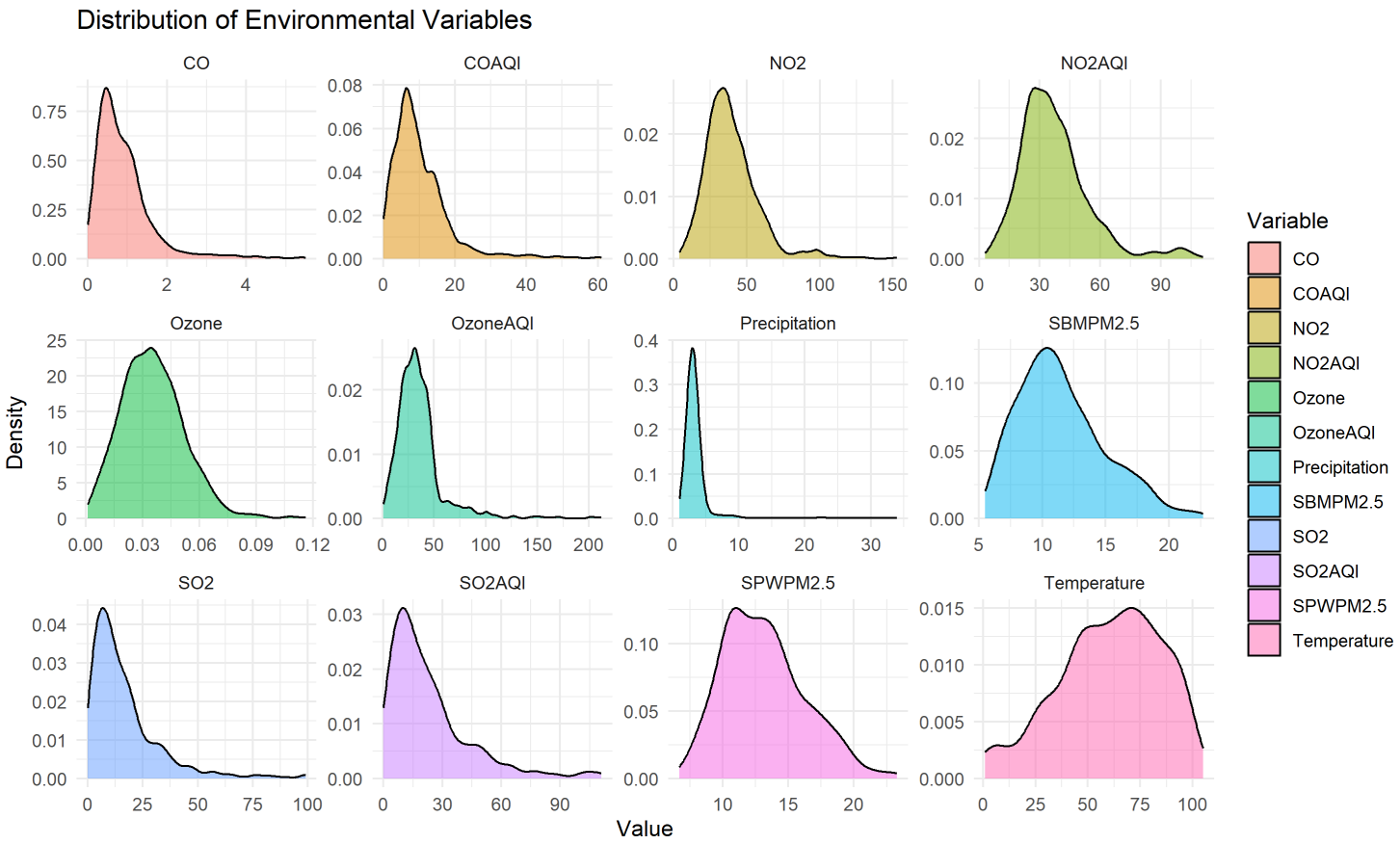
Distribution of Environmental Variables
CO
COAQI
NO2
NO2AQI
Ozone
OzoneAQI
Precipitation
SBMPM2.5
SO2
SO2AQI
SPWPM2.5
Temperature
Density
Value
Variable
CO
COAQI
NO2
NO2AQI
Ozone
OzoneAQI
Precipitation
SBMPM2.5
SO2
SO2AQI
SPWPM2.5
Temperature

Figure 1

Table 1: Statistical Summary of Variables (Median, Mean, and Range)

Variable	Median	Mean	Range
Precipitation	3.0	3.894	0.0 - 33.0
Temperature	62.5	60.45	0.0 - 104.0
SPWPM2.5	12.9	13.17	0.0 - 16.6
SBMPM2.5	10.95	11.52	0.0 - 17.2
CO	0.7	0.8995	0.0 - 5.5
COAQI	8.0	10.27	0.0 - 61.0
NO2	36.0	39.3	0.0 - 149.2
NO2AQI	34.0	37.22	0.0 - 108.0
Ozone	0.034	0.03547	0.0 - 0.115
OzoneAQI	31.0	35.76	0.0 - 211.0
SO2	12.0	16.83	0.0 - 99.0
SO2AQI	17.0	23.45	0.0 - 111.0
rset	110.5	117.94	0.0 - 524.0
white	1.0	0.745	0.0 - 1.0
educbas	5.0	4.905	0.0 - 5.0
visit	3.5	3.5	2.0 - 3.5
macsvisit	325.0	335.8	0.0 - 650.0
hivatvisit	0.0	0.49	0.0 - 1.0
aar	-0.5969	0.0	-51.8507 - 0.0
eeaa	-1.064	0.0	-35.742 - 0.0
peaa	-0.2639	0.0	-33.6462 - 0.0
geaa	-0.8503	0.0	-43.2309 - 0.0
dnamtladjage	-0.00233	0.0	-2.37413 - 0.0
lnabscd42	6.560	6.417	4.578 - 9.851
lnabscd82	6.546	6.554	5.050 - 8.531
cd4nadir	444.5	478.4	0.0 - 1384.0
lnabsnaivecd4	5.449	5.221	2.063 - 7.938
lnabsnaivecd8	-	-	-
lnabssencd4	-	-	-
lnabssencd8	-	-	-
lnabsactcd4	-	-	-
lnabsactcd8	-	-	-
lncd3	0.5145	0.5037	0.0 - 0.4757
lncd4	0.42259	0.38787	0.0 - 0.59958
lncd8	0.3473	0.3644	0.0 - 0.5306
lnnaivecd4	0.10379	0.11002	0.0 - 0.35957
lnnaivecd8	5.0529	4.9076	0.0 - 6.7675
lnsencd4	0.00952	0.01457	0.0 - 0.09652
lnsencd8	0.04724	0.05679	0.0 - 0.29498
lnactcd4	0.01054	0.01135	0.0 - 0.03974
lnactcd8	0.01242	0.03201	0.0 - 0.37994
hcstv	0.0	0.045	0.0 - 1.0
hepb	0.0	0.02381	0.0 - 1.0
bmi	24.8	25.94	17.1 - 46.8
cum pkyear	0.0	2.9638	0.0 - 44.6
age	45.0	43.18	22.0 - 76.0
female	1.0	0.436	0.0 - 1.0
black	1.0	0.200	0.0 - 1.0
hispanic	0.01	0.105	0.0 - 1.0

Table 2: Meteorological and Air Pollution Factors

Factor	Unit
Precipitation	mm (scaling factor of 10)
Temperature	Fahrenheit (scaling factor of 10)
Statewide Population-Weighted PM2.5	µg/m³
Statewide Geographic-Mean PM2.5	µg/m³
Daily Max 8-hour CO Concentration	ppm
AQI Value for CO Concentration	-
Daily Max 1-hour NO2 Concentration	ppb
AQI Value for NO2 Concentration	-
Daily Max 8-hour Ozone Concentration	ppm
AQI Value for Ozone Concentration	-
Daily Max 1-hour SO2 Concentration	ppb
AQI Value for SO2 Concentration	-

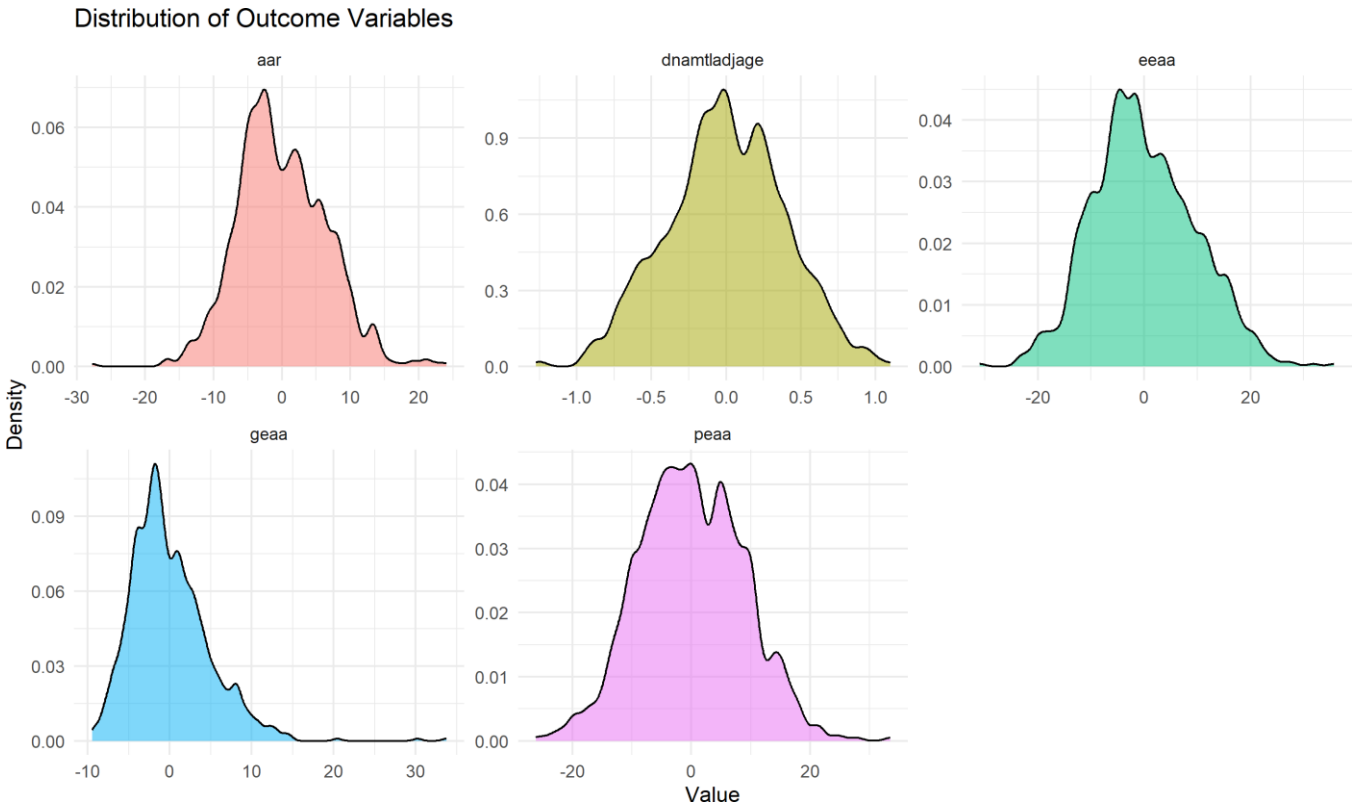

Figure 2

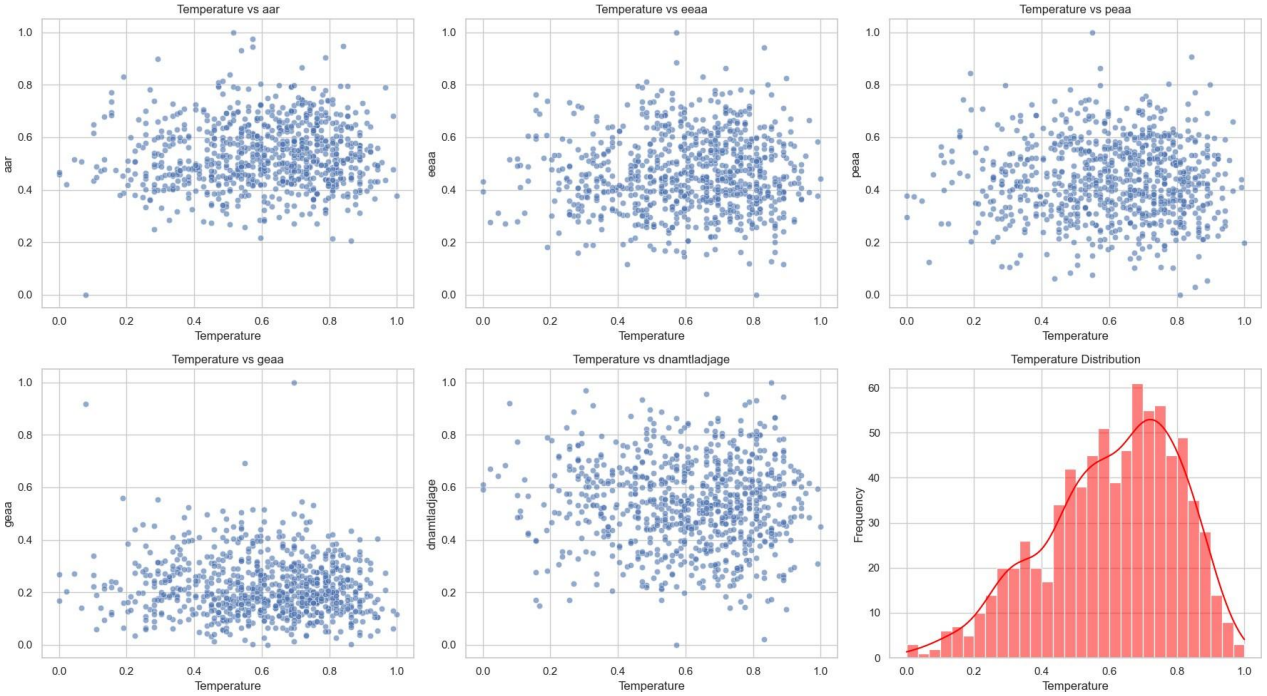

Figure 3

2.2

2.3 Methods

In our analysis, we first convert columns to numerical format. Next, we assess the proportion of missing values and employ imputation methods other than mean or median to avoid introducing excessive repeated values. Feature engineering will be conducted to create additional features, such as the square of all columns or interaction terms (e.g., x1 × x2). Through data integrations, we will begin with data visualization to examine the relationship between environmental factors and the epigenome. We construct features that refer to dependent variables, specifically interaction terms between environmental and epigenetic factors. Simultaneously, we believe that a single factor cannot diagnose or predict the outcome, so a combination of factors will be featured. We also recognize the existence of health inequality, since the data includes variables such as smoking, education, and BMI, we can further study health disparities. These analyses will provide profound insights into how socioeconomic factors influence the healthy aging process and offer theoretical and empirical support for future health policy formulation.

We introduce interaction terms to reveal the complex relationships between climate variables and regional characteristics, analyzing whether there are significant differences in the performance of dependent variables across different regions when faced with the same climatic conditions. In this analysis, macsidnumber, which is participant ID number, serves as a categorical variable representing different geographic locations, and its effect on the dependent variables reflects regional disparities.

3 Results

3.1 Data Visualization

To interpret our findings, we operationalized each climate and weather variable systematically. Scaled precipitation (ranging from 0.0 to 33.0 inches) and temperature (spanning 0.0 to 104.0°F) were included as key factors influencing health outcomes and epigenetic aging in the context of climate change. We analyzed their combined effects using advanced statistical methods. Additionally, statewide population-weighted PM2.5 and geographic-mean PM2.5 were used to assess long-term pollution exposure, while daily maximum concentrations of CO, NO2, ozone, and SO2 captured acute exposure risks. By integrating these environmental factors with biological metrics, this study provides valuable insights into how climate change and pollution disproportionately impact vulnerable populations, such as individuals living with HIV. These variables form a robust foundation for understanding environmental impacts on health.

We analyzed five key outcome variables: aar, eeaa, peaa, geaa, and dnamtladjage. The five measures were as follows: Age Acceleration Residual (AAR), derived from the residuals of a linear regression model between epigenetic age and chronological age [36]; Extrinsic Epigenetic Age Acceleration (EEAA), developed by Hannum [35], which reflects age-adjusted residuals positively associated with senescent T lymphocytes and negatively associated with naive T lymphocytes; Grim Epigenetic Age Acceleration (GEAA) [37], which predicts differences in both lifespan and healthspan; and Phenotypic Epigenetic Age Acceleration (PEAA) [38,39], which further assesses variations in biological aging and health outcomes.

The values of aar range from 0 to 1, with a mean of 0.49, indicating a noticeable skew in the distributions.

The variables eeaa, peaa, and geaa exhibit considerable variability, with minimum and maximum values of [−27.7784, 24.0723], [−30.931, 35.742], and [−26.1493, 33.4969], respectively. Notably, the medians of these variables are negative, suggesting that most observations are concentrated in the negative region. On the other hand, dnamtladjage ranges from −9.4753 to 33.7556, with a median of −0.8503, indicating a predominance of negative values. Overall, the statistical characteristics of these outcome variables reveal the complex relationships between health outcomes and influencing factors, providing crucial insights for understanding health-related risk factors.

The boxplot and violin plot for Grim Epigenetic Age Acceleration (GEAA) shows that under extreme precipitation conditions, the median value of GEAA is significantly lower compared to normal precipitation conditions. Since a lower GEAA typically suggests slower biological aging, this might seem like a positive outcome. However, the finding that extreme precipitation could be linked to such changes is unexpected and requires further investigation to understand the underlying mechanisms.

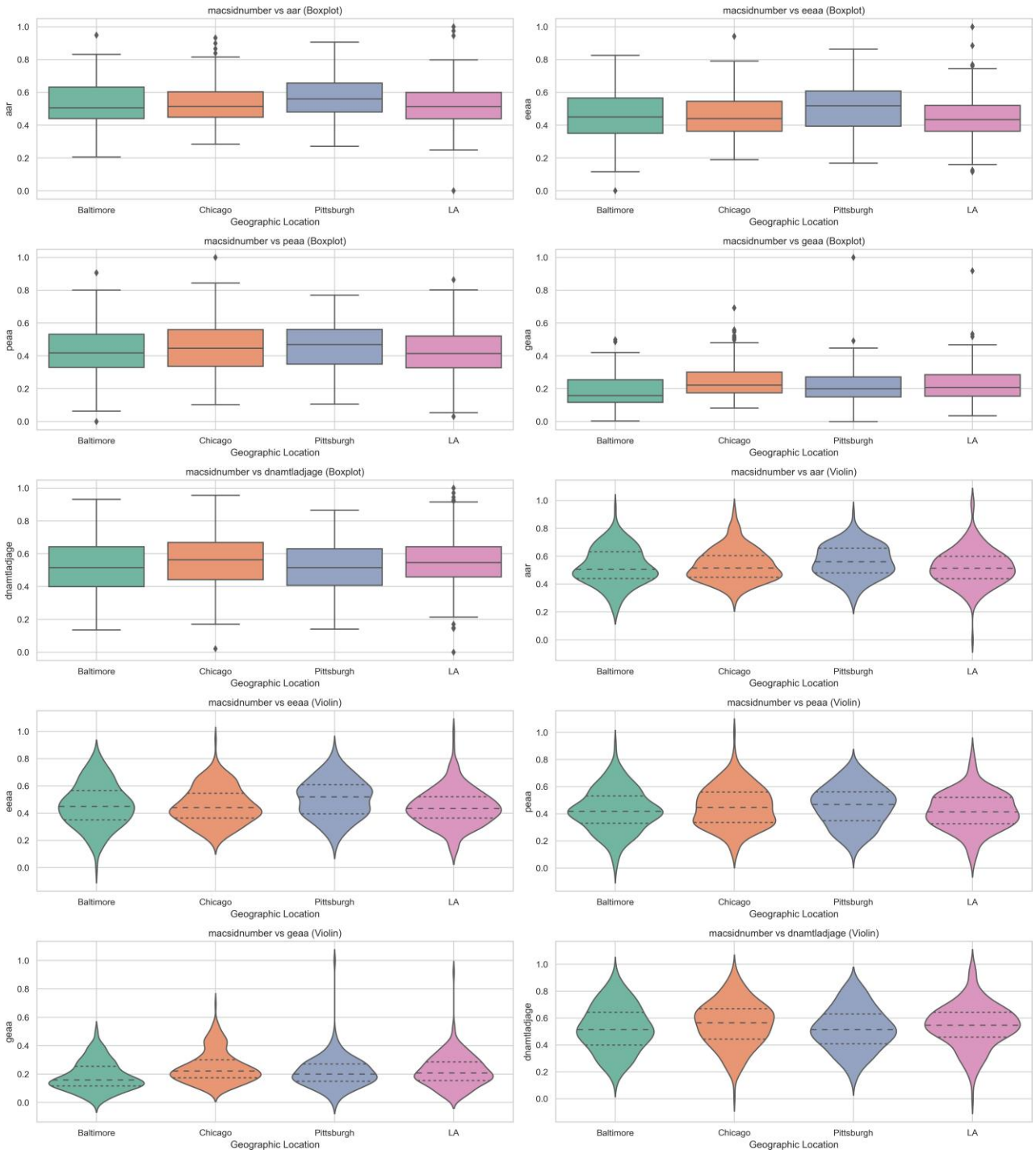
macsidnumber vs aar (Boxplot)
macsidnumber vs eeaa (Boxplot)
macsidnumber vs peaa (Boxplot)
macsidnumber vs geaa (Boxplot)
macsidnumber vs dnamtladjage (Boxplot)
macsidnumber vs aar (Violin)
macsidnumber vs eeaa (Violin)
macsidnumber vs peaa (Violin)
macsidnumber vs geaa (Violin)
macsidnumber vs dnamtladjage (Violin)
Baltimore
Chicago
Pittsburgh
LA
Geographic Location

Figure 4

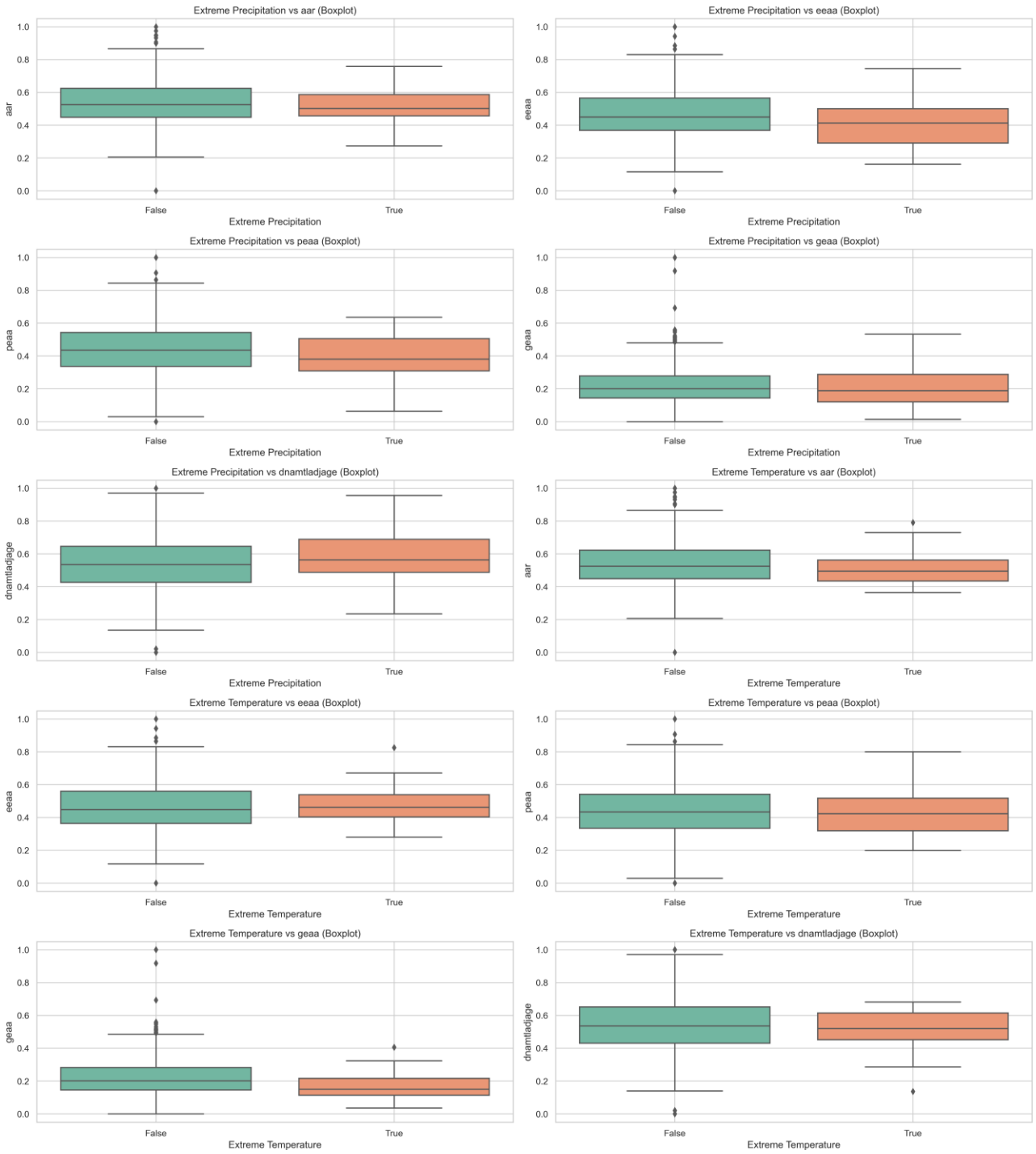
Extreme Precipitation vs aar (Boxplot)
aar
Extreme Precipitation
Extreme Precipitation vs eeaa (Boxplot)
eeaa
Extreme Precipitation
Extreme Precipitation vs peaa (Boxplot)
peaa
Extreme Precipitation
Extreme Precipitation vs geaa (Boxplot)
geaa
Extreme Precipitation
Extreme Precipitation vs dnamtladjage (Boxplot)
dnamtladjage
Extreme Precipitation
Extreme Temperature vs aar (Boxplot)
aar
Extreme Temperature
Extreme Temperature vs eeaa (Boxplot)
eeaa
Extreme Temperature
Extreme Temperature vs peaa (Boxplot)
peaa
Extreme Temperature
Extreme Temperature vs geaa (Boxplot)
geaa
Extreme Temperature
Extreme Temperature vs dnamtladjage (Boxplot)
dnamtladjage
Extreme Temperature
False
True

Figure 5

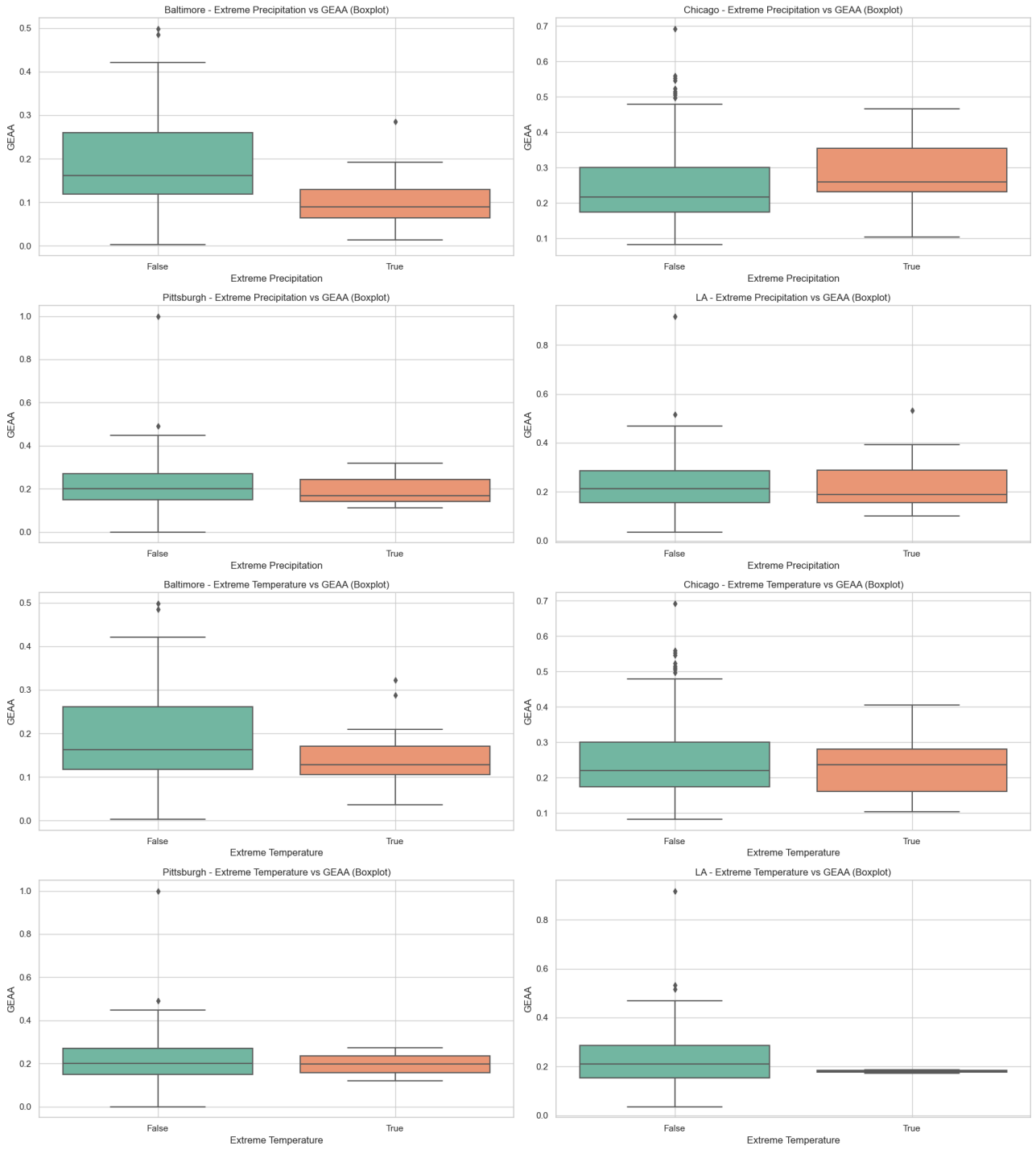
Baltimore - Extreme Precipitation vs GEAA (Boxplot)
Chicago - Extreme Precipitation vs GEAA (Boxplot)
Pittsburgh - Extreme Precipitation vs GEAA (Boxplot)
LA - Extreme Precipitation vs GEAA (Boxplot)
Baltimore - Extreme Temperature vs GEAA (Boxplot)
Chicago - Extreme Temperature vs GEAA (Boxplot)
Pittsburgh - Extreme Temperature vs GEAA (Boxplot)
LA - Extreme Temperature vs GEAA (Boxplot)
GEAA
False
True
Extreme Precipitation
Extreme Temperature

Figure 6

3.2 Analyzing Interaction Terms Between Climate, social demographic Variables and Regional Characteristics

We constructed regression models to explore the relationship between the dependent variable (such as 'aar,' 'eeaa,' etc.) and multiple independent variables, focusing particularly on the different cities represented by 'participant ID number,' and their interaction effects with climate factors (precipitation and temperature), racial composition ('white'), and educational level ('educbas').

The coefficients for 'Chicago' and 'Los Angeles' are significant ($p < 0.05$), indicating a positive correlation with the dependent variable, while the coefficient for 'Pittsburgh' is not significant ($p = 0.2225$), suggesting a weaker relationship. The climate factors, 'Precipitation' and 'Temperature,' show no significant effects ($p > 0.1$). Conversely, the 'white' coefficient is significant ($p < 0.001$), reflecting a negative impact where an increase in racial composition correlates with a decrease in the dependent variable. The 'educbas' coefficient is also significant ($p < 0.05$), indicating that higher education levels are associated with increased dependent variable values. Most interaction terms, such as 'Baltimore: Precipitation,' are not significant, while the interaction term 'Pittsburgh' is significant ($p = 0.0355$), suggesting that regional differences, possibly influenced by factors such as education, may play a role. The model's R-squared value of 0.1475 suggests that approximately 14.75% of the variation in the dependent variable is explained, indicating that other omitted variables may hold greater explanatory power. The overall model is significant, as shown by the F-statistic and its p-value (< 2.2e-16). In summary, the model reveals significant influences of regional variables, racial composition, and educational level on the dependent variable, while the effects of climate factors remain insignificant.

Prediction Performance

After standardizing the features, the dataset is divided into training, validation, and test sets. A function is then defined to construct a deep learning model, utilizing Keras Tuner for hyperparameter optimization to find the optimal network architecture. Once the best model is identified, it is trained and compared against a linear regression model. Finally, scatter plots in Figure 7 illustrate the relationship between actual and predicted values, enabling the assessment of both models' predictive performance. The optimal hyperparameters identified are: 32 units in the first layer and 16 units in the second layer.

We observed minimal differences in predictive performance between deep learning and linear regression models. Both approaches effectively captured the underlying patterns in the data, indicating that, given the dataset and the complexity of the relationships, the traditional linear regression model can be as effective as the more intricate deep learning method. This finding underscores the significance of selecting models based on the specific context and data characteristics rather than merely assuming that advanced techniques are superior. In cases where the underlying relationships are relatively straightforward, simpler models can provide competitive performances while being more interpretable and less computationally demanding.

Table 3: Model Summary

Variable	**Estimate**	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	0.512804	0.048217	10.635	<2e-16 ***
Chicago	0.139584	0.060411	2.311	0.0211 *
Los Angeles	0.146346	0.067583	2.165	0.0307 *
Pittsburgh	0.136616	0.111908	1.221	0.2225
Precipitation	0.151898	0.096080	1.581	0.1143
Temperature	0.066543	0.060010	1.109	0.2678
white	-0.136168	0.029192	-4.665	3.64e-06 ***
educbas	0.102539	0.043176	2.375	0.0178 *
Chicago:Precipitation	0.075838	0.157755	0.481	0.6308
Los Angeles:Precipitation	-0.217591	0.162489	-1.339	0.1809
Pittsburgh:Precipitation	-0.186892	0.177327	-1.054	0.2922
Chicago:Temperature	-0.131717	0.075437	-1.746	0.0812.
Los Angeles:Temperature	-0.090431	0.078252	-1.156	0.2482
Pittsburgh:Temperature	-0.096006	0.155303	-0.618	0.5366
Chicago:white	0.008985	0.037732	0.238	0.8119
Los Angeles:white	-0.023413	0.042427	-0.552	0.5812
Pittsburgh:white	-0.002310	0.040059	-0.058	0.9540
Chicago:educbas	-0.070922	0.059175	-1.199	0.2311
Los Angeles:educbas	-0.023455	0.062777	-0.374	0.7088
Pittsburgh:educbas	-0.132005	0.062689	-2.106	0.0355 *

323

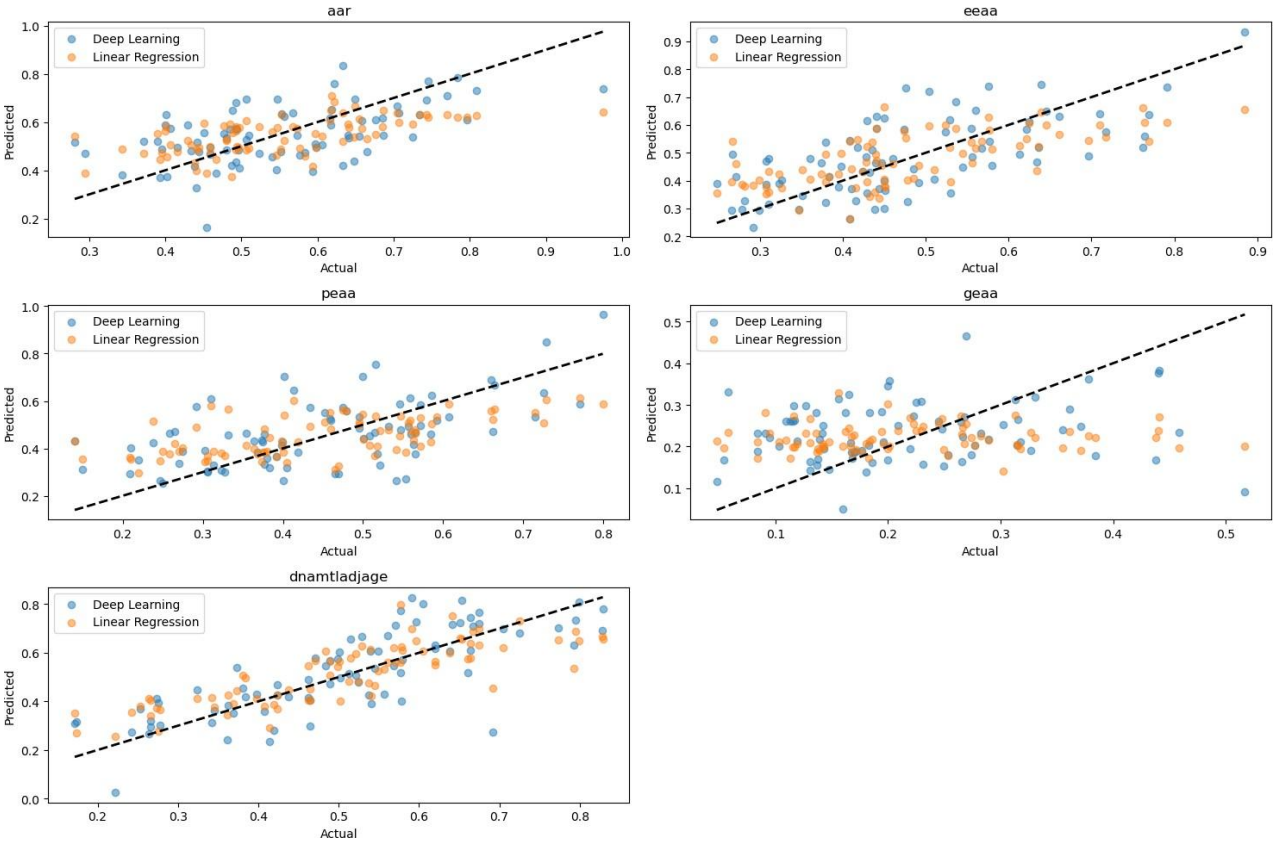

Figure 7

4 Conclusions

Our research investigated the effects of weather and air pollution variables, beyond heatwaves, on epigenetic aging, particularly in the context of people aging with HIV. By utilizing deep learning methods on the dataset from four U.S. cities, including Los Angeles, we aimed to provide insights into climate and public health policies that mitigate the impacts of climate change on individuals living with HIV.

We introduced interaction terms to investigate the intricate relationships between climate variables and regional characteristics, revealing significant disparities in health outcomes. Our regression analysis showed strong positive correlations for Chicago and Los Angeles with the dependent variables, while Pittsburgh exhibited a weaker association. Baltimore, on the other hand, did not show a significant correlation Importantly, climate factors like precipitation and temperature had minimal impact on epigenetic aging. In contrast, racial compositions demonstrated a significant negative correlation, suggesting that a higher percentage of white residents was linked to lower levels of epigenetic aging. Furthermore, education level positively affected these outcomes, particularly highlighted by the significant interaction between education and the Pittsburgh region.

The model explained approximately 14.75% of the variance in the dependent variables, suggesting that other unmeasured factors may play a crucial role. Despite the minimal influence of climate factors, the overall model was statistically significant.

Moreover, our analysis revealed minimal differences in predictive performance between deep learning and linear regression models, with both successfully capturing the underlying data patterns. This indicates that traditional linear regression can be just as effective as more complex deep learning methods, highlighting the significance of selecting models based on the specific context and characteristics of the data. In situations where the relationships are relatively straightforward, simpler models can deliver competitive performance while offering enhanced interpretability and lower computational demands.

Extreme precipitation can have complex effects on health outcomes, especially for PLWH. While it may not directly improve GEAA, it can lead to increased humidity, flooding, and associated health issues, such as respiratory problems or infections, which in turn may negatively affect overall health. Thus, while the result suggests a decrease in GEAA, it's important to consider the broader health implications of extreme weather events, which can complicate the relationship between extreme precipitation and health outcomes. Additionally, extreme temperatures may impact the physiological state of organisms and environmental conditions, exacerbating the effects of climate change and further influencing GEAA. Based on these findings, policymakers should consider measures to mitigate the impact of extreme weather events. we are looking at county-wide data of the four counties.

This pilot project utilizes previously collected data from the MACS. Due to this limitation, our study currently includes only male individuals, a significant portion of whom identify as white. However, our future studies will be expanded to include women enrolled in the WIHS, and if funding is secured, we could also increase the number of participants to better capture the differential effects of sex and race/ethnicity on the impact of weather on epigenetic aging in PLWH. Data that need to be considered in the future include socioeconomic variables, such as cognition, sleep, nutrition, diet, exercise, etc. With those data, we could analyze how factors such as cognition, sleep, nutrition, diet, and exercise affect epigenetic age and aging rates.

In conclusion, this study underscores the profound impact of regional variables, racial compositions, and education on epigenetic aging, while revealing the minimal influence of climate factors. These findings emphasize the importance of addressing health inequalities and advocate for nuanced policy-making that takes into account the intricate interplay between environmental and social variables in shaping health outcomes.

Data and code availability statement

Epigenetic data from the Multicenter AIDS Cohort Study (MACS) will be provided upon approval of the concept sheet submitted to the MWCCS (https://statepi.jhsph.edu/mwccs/) Please reference concept sheet C-15039. Air quality and climate data have been collected from the United States Environmental Protection Agency (EPA), the National Centers for Environmental Information (NCEI), and the Atmospheric Composition Analysis Group at Washington University in St. Louis. Data will available upon request.

The code is available here: https://github.com/congca/Health-Inequalities-Aging-in-HIV–Deep-Learning

Acknowledgments

Primary data in this manuscript were collected by the Multicenter AIDS Cohort Study (MACS), now the MACS/WIHS Combined Cohort Study (MWCCS), The contents of this publication are solely the responsibility of the authors and do not represent the official views of the National Institutes of Health (NIH). MWCCS (Principal Investigators): Baltimore CRS (Todd Brown and Joseph Margolick), U01-HL146201; Data Analysis and Coordination Center (Gypsyamber D'Souza, Stephen Gange and Elizabeth Topper), U01-HL146193; Chicago-Cook County CRS (Mardge Cohen, Audrey French, and Ryan Ross), U01-HL146245; Chicago-Northwestern CRS (Steven Wolinsky, Frank Palella, and Valentina Stosor), U01-HL146240; Los Angeles CRS (Roger Detels and Matthew Mimiaga), U01-HL146333; Pittsburgh CRS (Jeremy Martinson and Charles Rinaldo), U01-HL146208; The MWCCS is funded primarily by the National Heart, Lung, and Blood Institute (NHLBI), with additional co-funding from the Eunice Kennedy Shriver National Institute Of Child Health & Human Development (NICHD), National Institute On Aging (NIA), National Institute Of Dental & Craniofacial Research (NIDCR), National Institute Of Allergy And Infectious Diseases (NIAID), National Institute Of Neurological Disorders And Stroke (NINDS), National Institute Of Mental Health (NIMH), National Institute On Drug Abuse (NIDA), National Institute Of Nursing Research (NINR), National Cancer Institute (NCI), National Institute on Alcohol Abuse and Alcoholism (NIAAA), National Institute on Deafness and Other Communication Disorders (NIDCD), National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), National Institute on Minority Health and Health Disparities (NIMHD), and in coordination and alignment with the research priorities of the National Institutes of Health, Office of AIDS Research (OAR). MWCCS data collection is also supported by UL1-TR000004 (UCSF CTSA), UL1-TR003098 (JHU ICTR), UL1-TR001881 (UCLA CTSI), P30-AI-050409 (Atlanta CFAR), P30-AI-073961 (Miami CFAR), P30-AI-050410 (UNC CFAR), P30-AI-027767 (UAB CFAR), P30-MH-116867 (Miami CHARM), UL1-TR001409 (DC CTSA), KL2-TR001432 (DC CTSA), and TL1-TR001431 (DC CTSA).

The authors gratefully acknowledge the contributions of the study participants and dedication of the staff at the MWCCS sites.

CC and RMA acknowledge funding from the Caltech Linde Center for Science, Society, and Public Policy. B.D. Jamieson is also supported by U01-HL146333, and M.E. Sehl is the recipient of the Susan G. Komen Career Catalyst Award CCR16380478. C. Ramirez receives support from P30 MH058107. Will add more.

Conflicts of interest

The authors declare no conflict of interest.

References

[1] Breen, E. C. et al. Accelerated aging with HIV occurs at the time of initial HIV infection. iScience 25 (2022).

[2] Zhang, J. et al. Effects of highly active antiretroviral therapy initiation on epigenomic DNA methylation in persons living with HIV. Frontiers in Bioinformatics 4, 1357889 (2024).

[3] Olatosi, B. et al. Application of machine-learning techniques in classification of HIV medical care status for people living with HIV in South Carolina. AIDS (London, England) 35, S19–S28 (2021).

[4] Saldana, C. et al. Development of a machine learning modelling tool for predicting HIV incidence using public health data from a county in the Southern United States. Clinical Infectious Diseases: An Official Publication of the Infectious Diseases Society of America (2024).

[5] Orel, E. et al. Prediction of HIV status based on socio-behavioural characteristics in East and Southern Africa. PLOS ONE 17, 1–15 (2022). URL https://doi.org/10.1371/journal.pone.0264429.

[6] Conn, D., Ramirez, C. M. & Alvarez, R. Random forests and fuzzy forests in biomedical research. (2016).

[7] Conn, D., Ngun, T., Li, G. & Ramirez, C. M. Fuzzy forests: Extending random forest feature selection for correlated high-dimensional data. Journal of Statistical Software 91, 1–25 (2019).

[8] Watts, N. et al. The 2020 report of the Lancet countdown on health and climate change: Responding to converging crises. The Lancet 397, 129–170 (2021).

[9] Moore, G. & Semple, J. L. The impact of global warming on Mount Everest. High Altitude Medicine & Biology 10, 383–385 (2009).

[10] O'Gorman, P. A. Precipitation extremes under climate change. Current Climate Change Reports 1, 49–59 (2015).

[11] Eichelberger, S., McCaa, J., Nijssen, B. & Wood, A. Climate change effects on wind speed. North American Windpower 7, 68–72 (2008).

[12] Sun, D., Ji, C., Sun, W., Yang, Y. & Wang, H. Accuracy assessment of three remote sensing shortwave radiation products in the Arctic. Atmospheric Research 212, 296–308 (2018).

[13] Liang, W. et al. Long-term exposure to ambient particulate matter is associated with prognosis in people living with HIV/AIDS: Evidence from a longitudinal study. Science of The Total Environment 928, 172453 (2024). URL https://www.sciencedirect.com/science/article/pii/S0048969724025993.

[14] Zhang, F. et al. Ambient particulate matter, a novel factor hindering life spans of HIV/AIDS patients: Evidence from a ten-year cohort study in Hubei, China. Science of The Total Environment 875, 162589 (2023). URL https://www.sciencedirect.com/science/article/pii/S0048969723012056.

[15] Chen, H. et al. Associations of ambient ozone exposure and CD4+ T cell levels with mortality among people living with HIV: An eight-year longitudinal study. Science of The Total Environment 923, 171544 (2024). URL https://www.sciencedirect.com/science/article/pii/S0048969724016851.

[16] Baranczuk, Z. et al. Socio-behavioural characteristics and HIV: Findings from a graphical modelling analysis of 29 Sub-Saharan African countries. Journal of the International AIDS Society 22, e25437 (2019).

[17] He, N. et al. Understanding medical distrust among African American/Black and Latino persons living with HIV with sub-optimal engagement along the HIV care continuum: A machine learning approach. Sage Open 11, 21582440211061314 (2021). URL https://doi.org/10.1177/21582440211061314.

[18] Vidrine, D. J. Cigarette smoking and HIV/AIDS: Health implications, smoker characteristics and cessation strategies. AIDS Education and Prevention 21, 3–13 (2009). URL https://doi.org/10.1521/aeap.2009.21.3_supp.3.

[19] Saito, A., Karama, M. & Kamiya, Y. HIV infection, and overweight and hypertension: A cross-sectional study of HIV-infected adults in Western Kenya. Tropical Medicine and Health 48 (2020).

[20] Ordóñez, C. & Marconi, V. Understanding HIV risk behavior from a sociocultural perspective. Journal of AIDS & Clinical Research 3 (2012).

[21] Wu, H. A deep learning-based hybrid feature selection approach for cancer diagnosis. Journal of Physics: Conference Series 1848, 012019 (2021). URL https://dx.doi.org/10.1088/1742-6596/1848/1/012019.

[22] Skaathun, B. et al. Interplay between Geography and HIV Transmission Clusters in Los Angeles County. Open Forum Infectious Diseases 8, ofab211 (2021). URL https://doi.org/10.1093/ofid/ofab211.

[23] Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular Cell 49, 359–367 (2013).

[24] Horvath, S. et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. Aging (Albany NY) 10, 1758 (2018).

[25] Lu, A. T. et al. DNA methylation-based estimator of telomere length. Aging (Albany NY) 11, 5895 (2019).

[26] Levine, M. E. et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY) 10, 573 (2018).

[27] Horvath, S. & Levine, A. J. HIV-1 infection accelerates age according to the epigenetic clock. The Journal of Infectious Diseases 212, 1563–1573 (2015).

[28] Sehl, M. E. et al. The effects of anti-retroviral therapy on epigenetic age acceleration observed in HIV-1-infected adults. Pathogens and Immunity 5, 291 (2020).

[29] Sehl, M. E. et al. Increased rate of epigenetic aging in men living with HIV prior to treatment. Frontiers in Genetics 12, 796547 (2022).

[30] Weipeng Xing, G. Z. & Poslad, S. Estimation of global horizontal irradiance in China using a deep learning method. International Journal of Remote Sensing 42, 3899–3917 (2021). URL https://doi.org/10.1080/01431161.2021.1887539.

[31] Everson FP. HIV/AIDS and air pollution as emerging cardiovascular risk factors in Cape Town populations: is endothelial function a marker of effect (Doctoral dissertation, Stellenbosch: Stellenbosch University).

[32] Zhu S, Zhang F, Xie X, Zhu W, Tang H, Zhao D, Ruan L, Li D. Association between long-term

exposure to fine particulate matter and its chemical constituents and premature death in individuals living with HIV/AIDS. Environmental Pollution. 2024 Jun 15;351:124052.
[33] Padhi BK, Khatib MN, Ballal S, Bansal P, Bhopte K, Gaidhane AM, Tomar BS, Ashraf A, Kumar MR, Chauhan AS, Sah S. Association of exposure to air pollutants and risk of mortality among people living with HIV: a systematic review. BMC Public Health. 2024 Nov 22;24(1):3251.
[34] Zhang F, Tang H, Zhao D, Zhu S, Ruan L, Zhu W. Short-term exposure to ozone and mortality from AIDS-related diseases: A case-crossover study in the middle Yangtze River region, China. Preventive Medicine. 2023 Oct 1;175:107689.
[35] Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. Molecular Cell 49, 359–367 (2013).
[36] Horvath, S. et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. Aging (Albany NY) 10, 1758 (2018).
[37] Lu, A. T. et al. DNA methylation-based estimator of telomere length. Aging (Albany NY) 11, 5895 (2019).
[38] Levine, M. E. et al. An epigenetic biomarker of aging for lifespan and healthspan. Aging (Albany NY) 10, 573 (2018).
[39] Horvath, S. & Levine, A. J. HIV-1 infection accelerates age according to the epigenetic clock. The Journal of Infectious Diseases 212, 1563–1573 (2015).

Confidentiality protection and scientific rationale

This study is associated with Concept Sheet X24047. To protect the confidentiality of subject information in this study, we will ensure that all data received and used is de-identified. This involves removing any personal identifiers from the datasets, ensuring that individual subjects cannot be linked to the data. Additionally, any findings or reports generated will not disclose individual-level data.

The proposed study aims to examine the interplay between climate change, air pollution, and epigenetic aging, specifically in people living with HIV (PLWH). By integrating high-resolution meteorological and air pollution data with existing genome-wide epigenetic datasets, the research will utilize advanced machine-learning techniques to explore how environmental factors contribute to epigenetic aging. The study addresses gaps in existing research, which often suffers from small sample sizes and high dimensionality, by applying machine learning to manage complex datasets and identify key influences. The results will enhance our understanding of how climate change and environmental conditions impact aging processes, particularly in vulnerable populations, and will inform targeted public policy recommendations to mitigate health risks associated with climate change.

Author Contribution Statement

RMA, BJ, CR, and CC conceptualized the research.

CC developed the research question hypothesis, research design, methodology, and data analysis,

interpreted the results, and wrote the manuscript. CC and RH also conducted the literature review.

BJ and CR edited it and supported data acquisition and medical knowledge.

RMA, BJ, and CR reviewed and edited it. RMA secured funding for the project and CR provide the overall Supervision.

SWB guided the concept of the research.

FP, MM, JM, TB, SG, and SWB contributed to the initial medical data collection.

All authors have read and agreed to the published version of the manuscript.

Software environment

The Python environment used is Python 3.11.5, packaged by Anaconda, Inc. (main, Sep 11, 2023, 13:26:23) [MSC v.1916 64-bit (AMD64)]. The R version used is R 4.3.1 (2023-06-16 ucrt)