Using causal deep learning to explore the roles of climate change, air pollution, and epigenetics in the aging of people with HIV

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

Preexisting data has been collected from men living with HIV and men living without HIV providing critical information on the relative rates of epigenetic aging in both groups of men. Combining that data with historical data on air pollution and weather in areas where those people live, we apply random forest and K-means clustering for feature selection and DoWhy and XGBoost those two algorithms for causal inference. We confirmed that increases in CD4 and CD8 T-cell counts, specifically active and senescent cells, are positively correlated between these cell types and increased epigenetic aging. Increases in certain air quality concentrations seem to have a slight positive correlation on epigenetic age acceleration as measured by the Grim clock. These observations suggest that epigenetics may play a significant role in HIV-accelerated aging, while also indicating that specific external factors, such as air quality and smoking, could have a marginal impact on this process

Keywords

Causal deep learning, HIV, aging, climate, air pollution

1. Introduction

The Human Immunodeficiency Virus (HIV) remains a significant global public health challenge, especially due to accelerated aging observed in people living with HIV (PLWH) compared to their counterparts without HIV (PLWoH) [1, 2]. Although life expectancy for PLWH has improved in recent years, there is limited research on how socio-demographic factors, climate change, and air pollution contribute to the accelerated aging process in this population. Previous studies, such as those by Merzouki et al. (2020) and Mutai et al. (2023), have demonstrated the effectiveness of machine learning in identifying subgroups of individuals at varying risk levels for HIV/AIDS, highlighting the potential for targeted interventions [3, 4]. Additionally, Olatosi et al. (2021) emphasized the role of machine learning in predicting HIV care outcomes, while Orel et al. (2022) showcased its application in high-yield HIV testing [5, 12]. Despite this progress, research on the specific mechanisms linking epigenetics, environmental factors, and HIV-related aging remains limited.

Epigenetics refers to the chemical modifications, such as the addition to, or subtraction of methyl groups to DNA that influence gene expression without altering the genetic sequence itself. Methylation changes have been documented to accumulate over time, providing an accurate measure of biological age (Steve Horvath et al., 2013), also referred to as epigenetic age. Understanding epigenetic aging is crucial because it reveals how infectious agents, socioeconomic lifestyle, and environmental factors, such as air pollution and climate change, can accelerate aging, especially in vulnerable populations like PLWH. By using epigenetic aging markers, such as Horvath's Pan-tissue Age and Grim Age, we can gain insights into the biological aging process in PLWH (Steve Horvath et al., 2013). This allows for more precise assessments of health risks and can inform public health interventions aimed at reducing the impact of accelerated aging on this population.

In this study, we hypothesize that PM2.5 exposure, temperature, and precipitation interact to accelerate epigenetic aging in individuals living with HIV. Research has shown that PM2.5 exposure promotes aging by inducing inflammation, oxidative stress, and epigenetic changes, particularly through accelerated telomere shortening and cellular aging. Extreme temperatures and variations in precipitation may exacerbate these processes, as high temperatures can increase ground-level ozone and other pollutants, while lower precipitation levels lead to higher PM2.5 concentrations, intensifying oxidative stress and inflammation. In contrast, precipitation may help mitigate these effects by clearing pollutants from the air, potentially reducing the acceleration of aging. Using advanced machine learning techniques, we aim to explore the relationship between epigenetic aging and environmental factors such as air pollution, temperature, and precipitation, particularly in individuals living with HIV, to better understand how climate change may influence aging through air pollution. In this study, we use advanced machine learning techniques to explore the relationship between epigenetic aging and environmental factors such as air pollution and climate change, for example, precipitation, and temperature, leveraging data from men from the MACS/WIHS Combined Cohort Study (MWCCS) living with and at risk for HIV in the United States. We only focus on white males, because the males in the MWCCS cohort have been predominately white MSM. To identify key factors affecting HIV-related aging, we employ clustering methods (K-means) and supervised learning techniques (random forests). We further investigate causal relationships using gradient-boosting trees (XGBoost) and DoWhy, a framework designed for causal inference. This allows us to examine the impact of various factors such as smoking, education, and environmental exposures on epigenetic aging. By integrating epigenetic data, environmental factors, and socio-demographic variables, our research aims to provide actionable insights for public policy, ultimately contributing to improved health outcomes for PLWH.

2. Data and methods

Study Participants:

We examined epigenetic aging in 800 men, from 1984 to 2015, focusing on five specific measures of biological aging shortly after HIV infection. The MACS/WIHS Combined Cohort Study (MWCCS) follows men and women living with or at risk for HIV. Participants provide health information during regular visits and donate blood and saliva for testing. The age range for this analysis spans from 0 to the upper limit specified for each variable, which is as follows: 51.8507 units for Age acceleration residual (AAR) [13], 35.742 units for Extrinsic epigenetic age acceleration (EEAA) [12], 33.6462 units for phenotypic epigenetic age acceleration (PEAA) [15,16], 43.2309 units for Grim epigenetic age acceleration (GEAA) [14], and 2.37413 units for epigenetic estimator of telomere length (DNAmTLADJAGE) [15,16]. We selected our participants based on the cohort used in the Breen et al. (2022) study, which focused on individuals living with HIV. Specifically, we utilized the post-seroconversion data from the participants in that study, which provided a unique opportunity to investigate epigenetic aging following HIV infection.

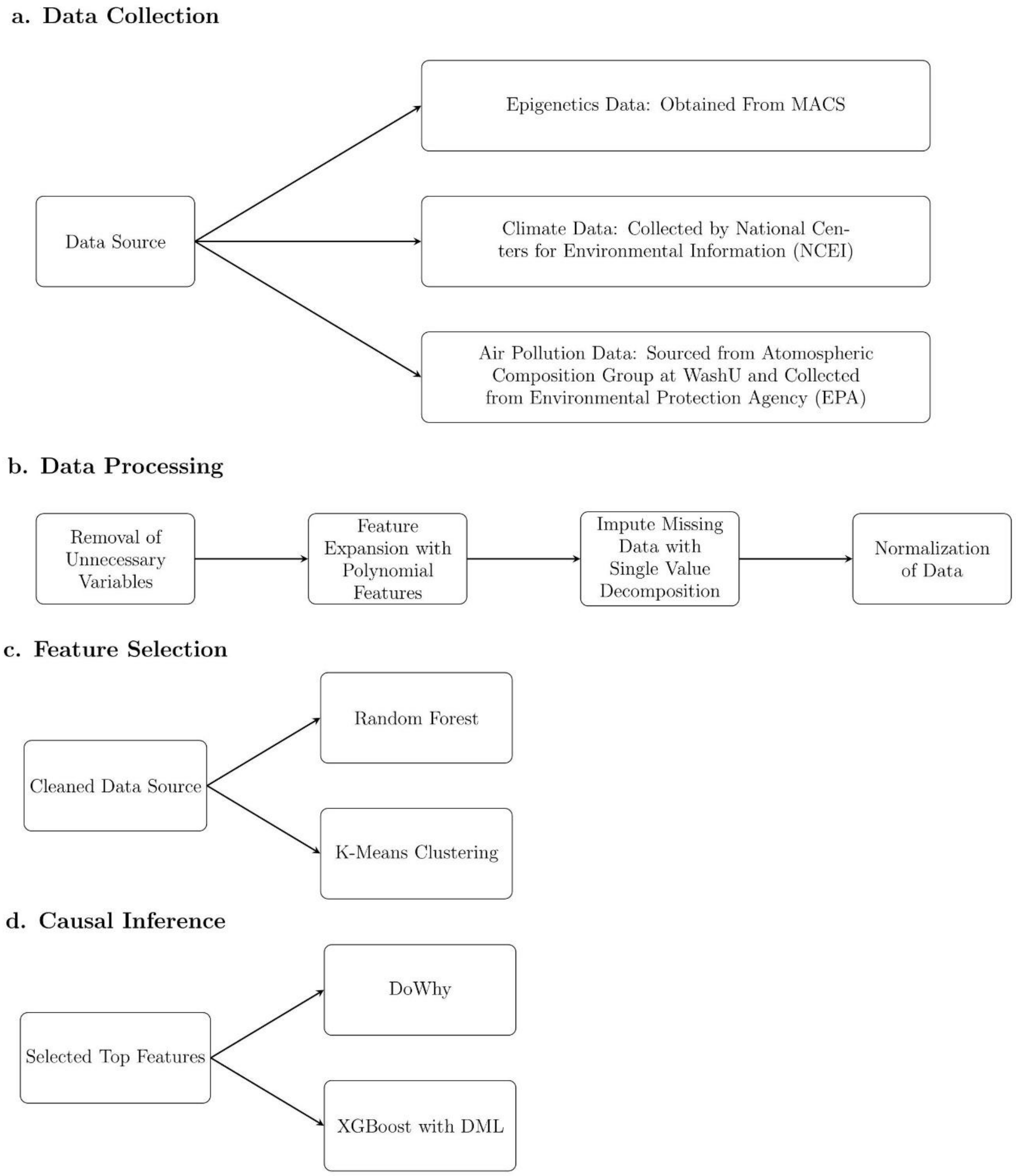
This approach allowed us to explore the relationship between HIV and aging by analyzing data collected after the participants had seroconverted, ensuring a focus on the impact of HIV on biological aging processes over time. We chose to follow a similar selection process because it enabled us to build upon their findings while leveraging robust longitudinal data to understand how HIV infection influences aging, particularly in the context of environmental factors like air pollution and climate change.

We matched participants on Hepatitis C Virus (HCV) status because co-infection with HCV can also impact biological aging and may confound the relationship between HIV and aging. Matching on HCV status ensured that we were isolating the effects of HIV on aging, rather than those driven by HCV co-infection, which is known to influence inflammation and immune system functioning. We limited our analysis to white men due to data availability and the need to ensure sufficient sample size for statistical analysis. We compared the PLWH group to a matched control group of men who did not contract HIV throughout the study period. These individuals were selected from the same pool of men at risk for HIV, ensuring that they had similar exposure to the same environmental and behavioral risk factors. We evaluated five specific epigenetic measures of biological aging over matched time intervals in both groups.

Overall Methodology:

Figure 1 illustrates the comprehensive approach we adopted to investigate the factors influencing epigenetic aging among men living with, and without, HIV. The methodology begins with data collection from men living with HIV and those without, encompassing socio-demographic information and historical environmental data such as air pollution, precipitation, and temperature. This data is then preprocessed to handle missing values and normalize variables. We utilize clustering techniques for feature selection, including random forests and K-means clustering, identifying the most relevant factors, including epigenetics and smoking habits. To explore causal relationships, we employ DoWhy and XGBoost, which enable us to assess the impacts of selected factors while controlling for confounding variables. Finally, we analyze the results to draw meaningful insights and develop evidence-based policy recommendations aimed at addressing the identified contributors to HIV aging.

Figure 1: Overall methodology used in the study


a. Data Collection
Data Source
Epigenetics Data: Obtained From MACS
Climate Data: Collected by National Centers for Environmental Information (NCEI)
Air Pollution Data: Sourced from Atomospheric Composition Group at WashU and Collected from Environmental Protection Agency (EPA)
b. Data Processing
Removal of Unnecessary Variables
Feature Expansion with Polynomial Features
Impute Missing Data with Single Value Decomposition
Normalization of Data
c. Feature Selection
Cleaned Data Source
Random Forest
K-Means Clustering
d. Causal Inference
Selected Top Features
DoWhy
XGBoost with DML

Data Collection and Processing

For each person, five measures of epigenetic aging were assessed. Although for this study we utilized epigenetic data from only the one-time point after the acquisition of HIV in the PLWH, all measures of epigenetic aging were calculated from raw data on each individual regressed across two timepoints, the pre-HIV acquisition as well as the post-HIV acquisition time-points as described by Breen et al. (2022). The five measures were as follows: Age acceleration residual (AAR) calculated from a linear regression model between the epigenetic age and chronologic age [13], Extrinsic epigenetic age acceleration (EEAA) [12], developed by Hannum, is another calculated age-adjusted residual that is positively correlated with senescent T lymphocytes and negatively correlated with naive T lymphocytes. Additional epigenetic measures were also assessed, predicting differences not only in health span but also lifespan such as Grim epigenetic age acceleration (GEAA) [14], and phenotypic epigenetic age acceleration (PEAA) [15,16]. Finally, an epigenetic estimator of telomere length (DNAmTLADJAGE) [15,16], an indicator of cellular proliferative history, was included as telomeres are known to shorten with age. To these measures of aging, additional information on these people was provided that may influence epigenetic aging, such as smoking, which greatly influences the Grim age clock and BMI.

To examine the relationship between climate change and air pollution on these measures of aging, additional data was compiled from various sources based on each person's location and visit date. For climate change variables, temperature and precipitation data were sourced from the National Centers for Environmental Information (NCEI). We focus on the monthly average temperature, rather than extreme heat. The temperature data is scaled, and the actual values need to be divided by 10 to obtain the temperature in Celsius. We choose to assess long-term climate patterns to understand the overall impact of temperature on health. We use the monthly cumulative precipitation, which is the total rainfall for all rainy days within the month. The precipitation data is also scaled, and the actual values need to be divided by 10 to get the amount in millimeters. We assess the impact of long-term precipitation patterns on health, rather than focusing on individual extreme precipitation events. Meanwhile, two different estimates of particulate matter less than 2.5 micrometers in diameter (PM2.5) were taken from the Atmospheric Composition Analysis Group at Washington University in St. Louis: population-weighted and geographic-mean PM2.5 estimates. Additional air quality concentrations, specifically carbon monoxide (CO), nitrogen dioxide (NO2), sulfur dioxide (SO2), and Ozone, and their respective AQI values were sourced from the Environmental Protection Agency (EPA).

After all additional data was added, the following steps were taken for data processing. First, data steps were taken to convert all data to raw data, as well as remove any unnecessary variables from the dataset. Then, feature expansion was performed by using polynomial features to expand the base set of roughly 50 features to over 200 features in hopes of potentially discovering any relationships between variables. Lastly, iterativeSVD (single value decomposition) was used to fill in any potential missing values contained within the dataset. The dates were converted to a numerical format, and min-max normalization was performed.

The case-control variable was recoded into binary values of 0 and 1. Interaction terms were created for all individual variables by computing pairwise products to examine the effect of these interactions on the outcome. This led to a dataset consisting of 2,092 variables and 800 observations.

Feature Selection

After data processing was complete, feature selection methods were applied to select the most relevant features for each of the five measures of epigenetic aging. The two methods used were random forest and K-means clustering. To visualize the results, Principal Component Analysis (PCA) was used to reduce the dimensionality of the data.

To select the final feature set for each of the epigenetic measures of aging, the top fifteen features from the random forest results were used in addition to the top ten closest relevant features for each of the measures of aging. With this, each target variable had a set of 25 features to be implemented for causal inference. More details are in Appendix A.

Causal Inference

We considered various factors, including climate, air pollution, socioeconomic demographics, and genetic variables, to understand their combined impact on health outcomes. These selected variables are directly input as features into algorithms or statistical analyses. We use causal deep learning as a key approach when exploring the effects of socioeconomic factors, such as smoking, education level, race, and BMI on healthy aging. The causal deep learning frameworks help us determine the true causal relationships between these factors and health outcomes, rather than merely correlations. We employ structured causal inference methods, including causal graphs, to clarify causal paths and potential covariates between variables, thereby reducing bias in causal inference. This approach enables us to handle complex nonlinear relationships and interactions, allowing for more accurate estimation and inference of causal effects.

Two main methods were used for causal inference: DoWhy and XGBoost. DoWhy is a Python library that takes advantage of two powerful frameworks for causal inference: graphical causal models and potential outcomes. In this study, we apply DoWhy to estimate the average treatment effect (ATE) and conditional ATE (CATE). Setting our case/control group as treatment variables (PLWH as the case group and PLWoH as the control group), we calculate the effect of HIV status on each of our measures of epigenetic aging (target variables) conditioned on each of our other features (covariates). Before computing the CATE, we first split each of the covariates into equal-width bins of percentiles. This allows us to study the impact on various percentile groups. Then, when computing the CATE, we perform a randomized intervention on the treatment variable in the fitted causal graph, draw samples from the interventional distribution, group observations by percentile bin, and then compute the treatment effect in each bin. After, we plot the corresponding confidence intervals.

XGBoost is a gradient-boosting framework that is commonly used for regression and prediction. To apply XGBoost in a causal inference setting, we use XGBoost in conjunction with DoubleML. Rather than calculating the CATE, we instead calculate the marginal effect of the treatment (MTE) on the outcome, conditioned on each covariate. We plot the marginal treatment effects for each epigenetic measure of aging and best-fit lines for each subplot for each feature.

3. Results

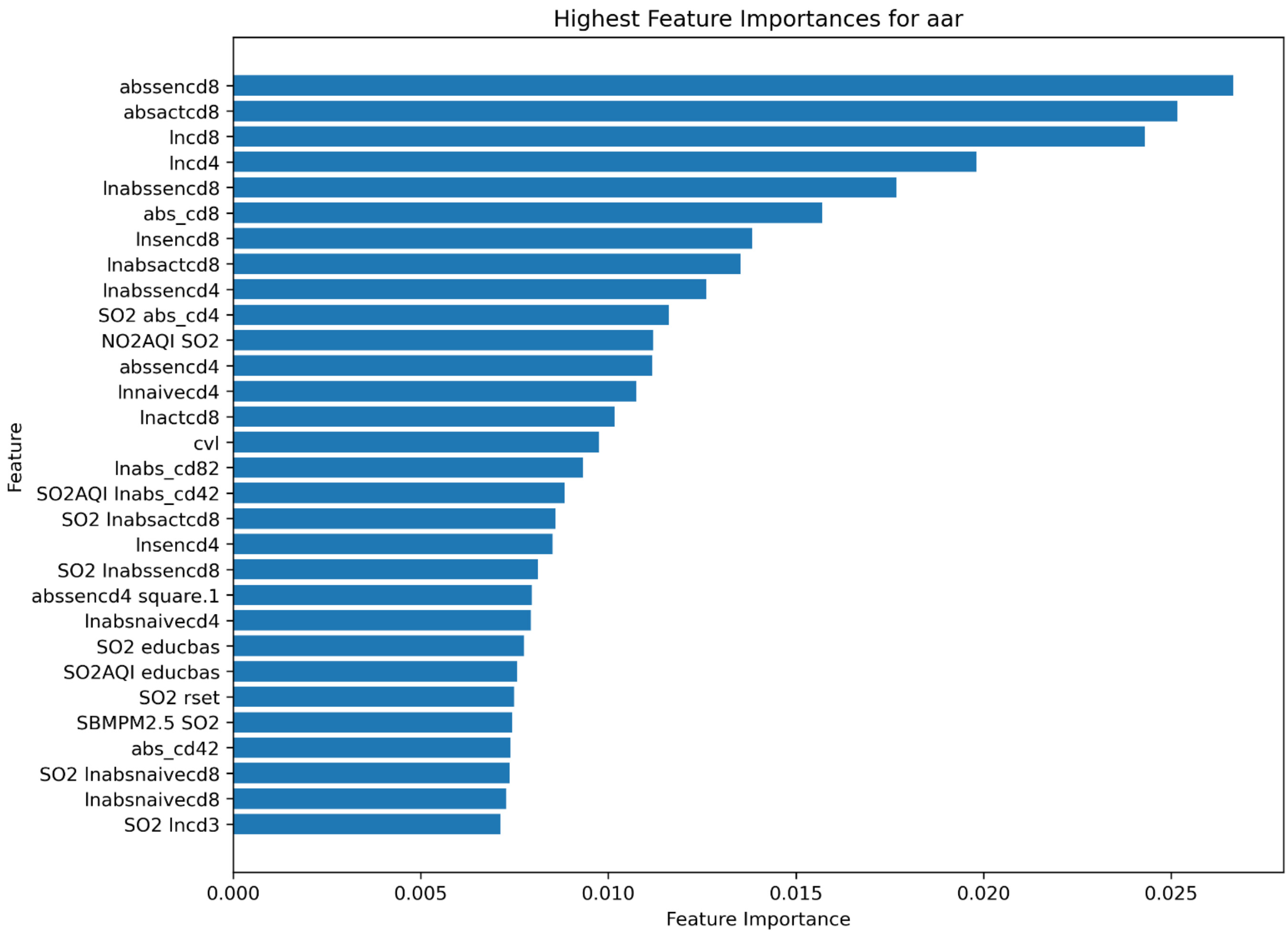
The results of this study highlight the role of epigenetic markers in measuring aging, particularly in individuals living with HIV. We found that HIV infection accelerated epigenetic aging, as evidenced by specific changes in DNA methylation patterns, which were more pronounced in individuals with higher inflammation and oxidative stress. Environmental factors such as air pollution (PM2.5) and temperature also influenced epigenetic aging, with higher PM2.5 exposure correlating with accelerated aging, while the effects of temperature were more variable. Participants with concurrent HCV infection showed more significant epigenetic changes compared to those without HCV, suggesting a compounded effect on aging. Additionally, when comparing HIV-positive individuals to HIV-negative individuals, we observed greater epigenetic aging in the HIV-positive group, despite matching for age and HCV status. The study found that **T-cell counts**, especially active and senescent T-cells, were closely related to several aging markers such as AAR, EEAA, PEAA, and DNAmTLadjage. However, the relationship is not straightforward, as an increase in T-cell counts does not necessarily lead to faster aging. This suggests that the impact of T-cells on aging in HIV-positive individuals is complex and cannot be solely predicted by changes in T-cell numbers.

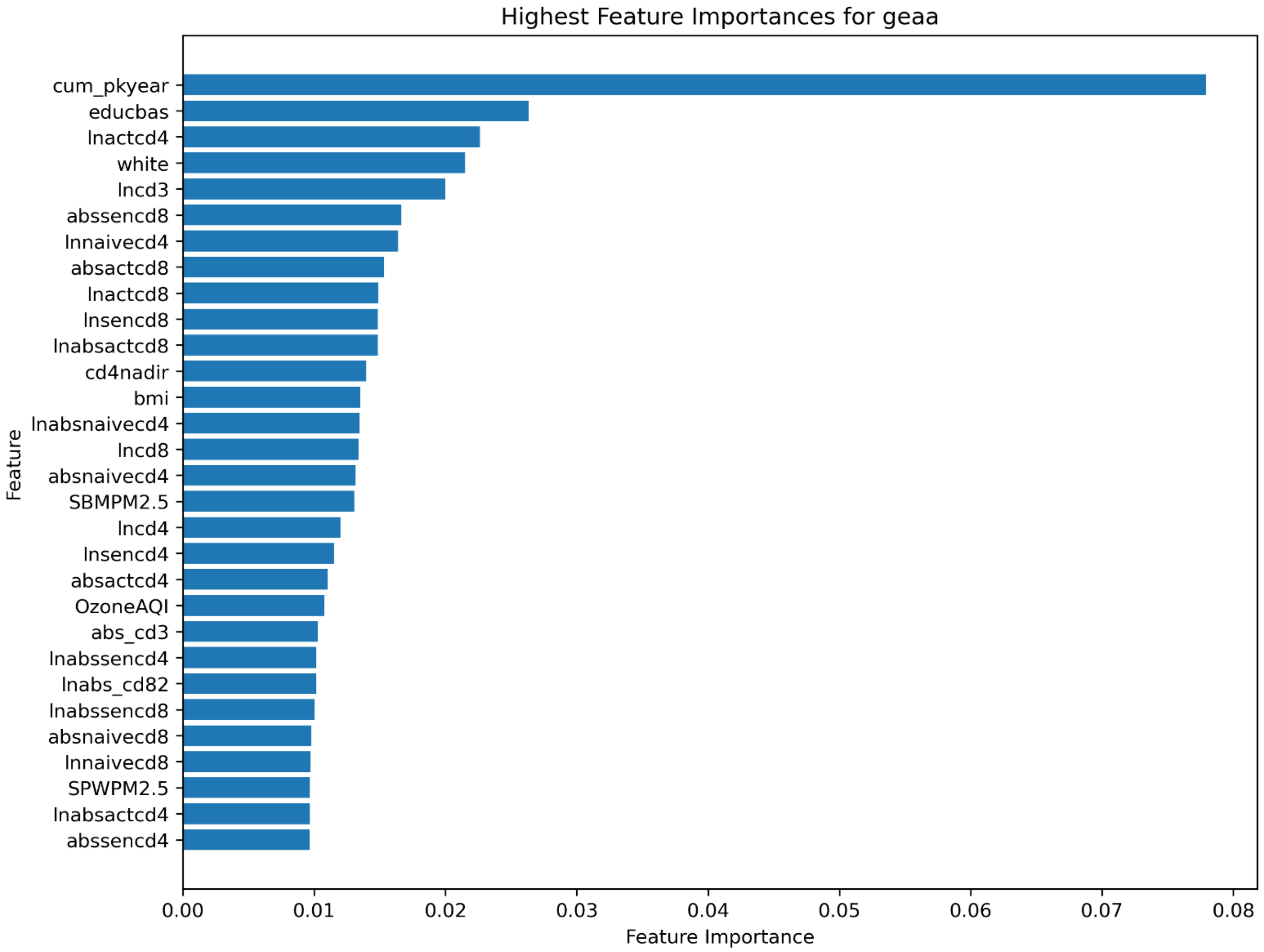
On the other hand, external factors also play a significant role in the aging process. The study found that for aging measures like AAR, EEAA, PEAA, and GEAA, air quality and climate were more closely linked to aging than biological markers. For example, an increase in ozone concentration was associated with accelerated aging in AAR and GEAA, indicating that air pollution has an impact on these aging markers. However, temperature changes had a relatively smaller effect on aging in individuals with HIV.

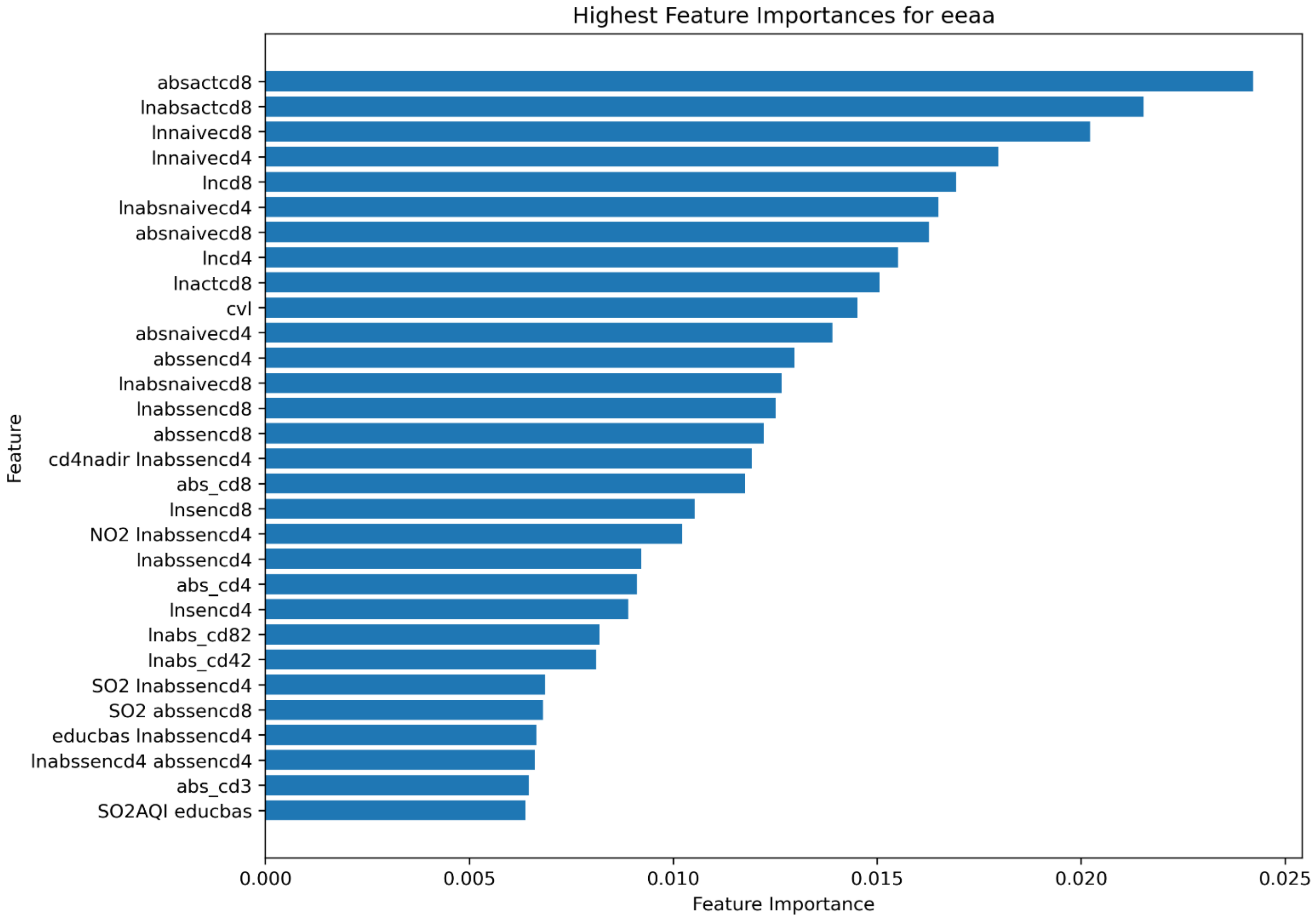
Specifically, for the aging measure, the study found that smoking was the most influential external factor, with higher smoking levels leading to faster aging. In contrast, individuals with higher levels of education showed slower aging in this measure. These findings suggest that in addition to biological markers, lifestyle habits, and social factors also play an important role in the aging process.

In summary, while epigenetic markers are critical in measuring aging, external factors like smoking, air quality, and education level also significantly influence the aging process, particularly in people living with HIV. The findings highlight the importance of environmental factors in aging, while also revealing the complex relationship between T-cell numbers and aging.

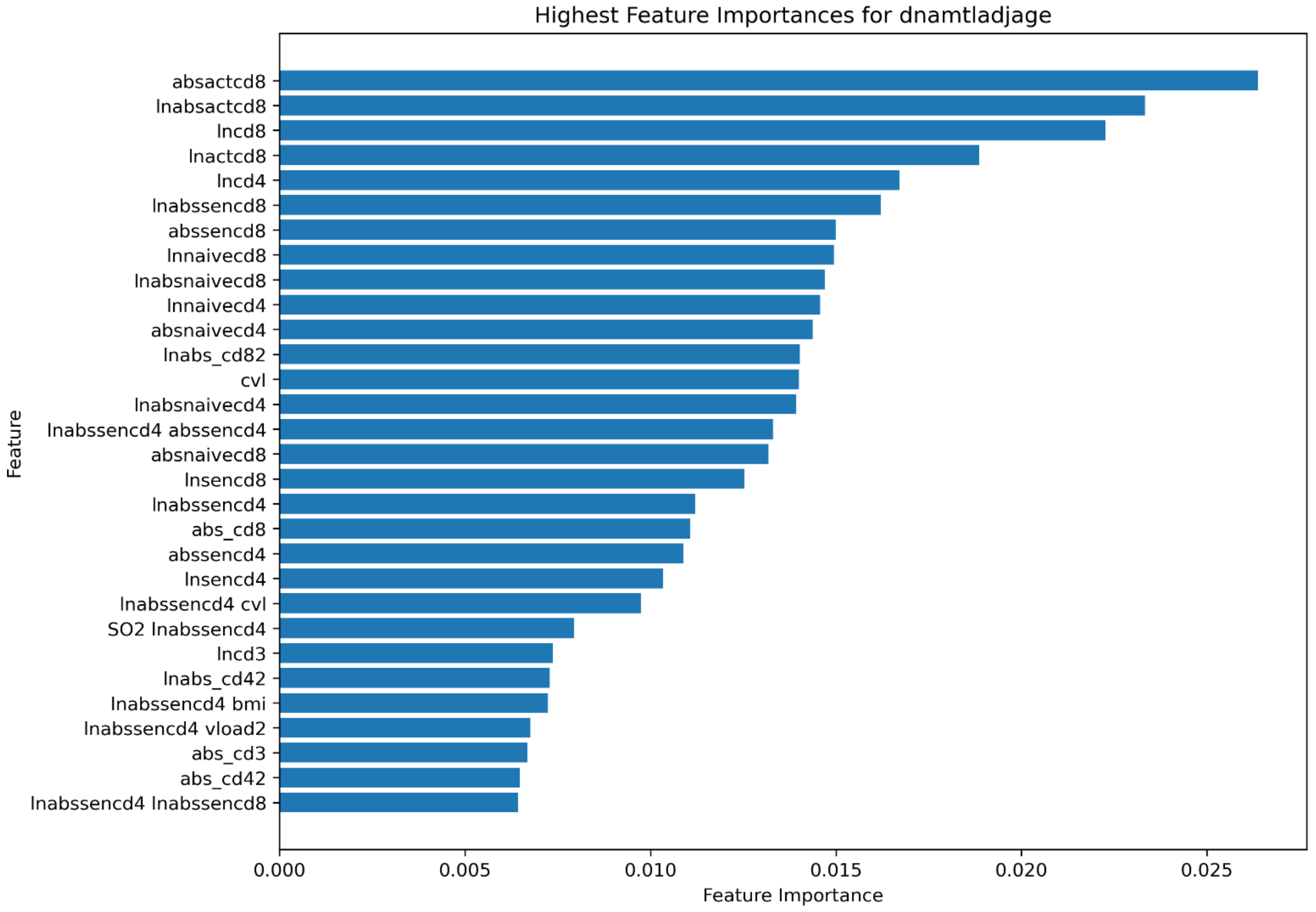
Figure 2: Random Forest Results



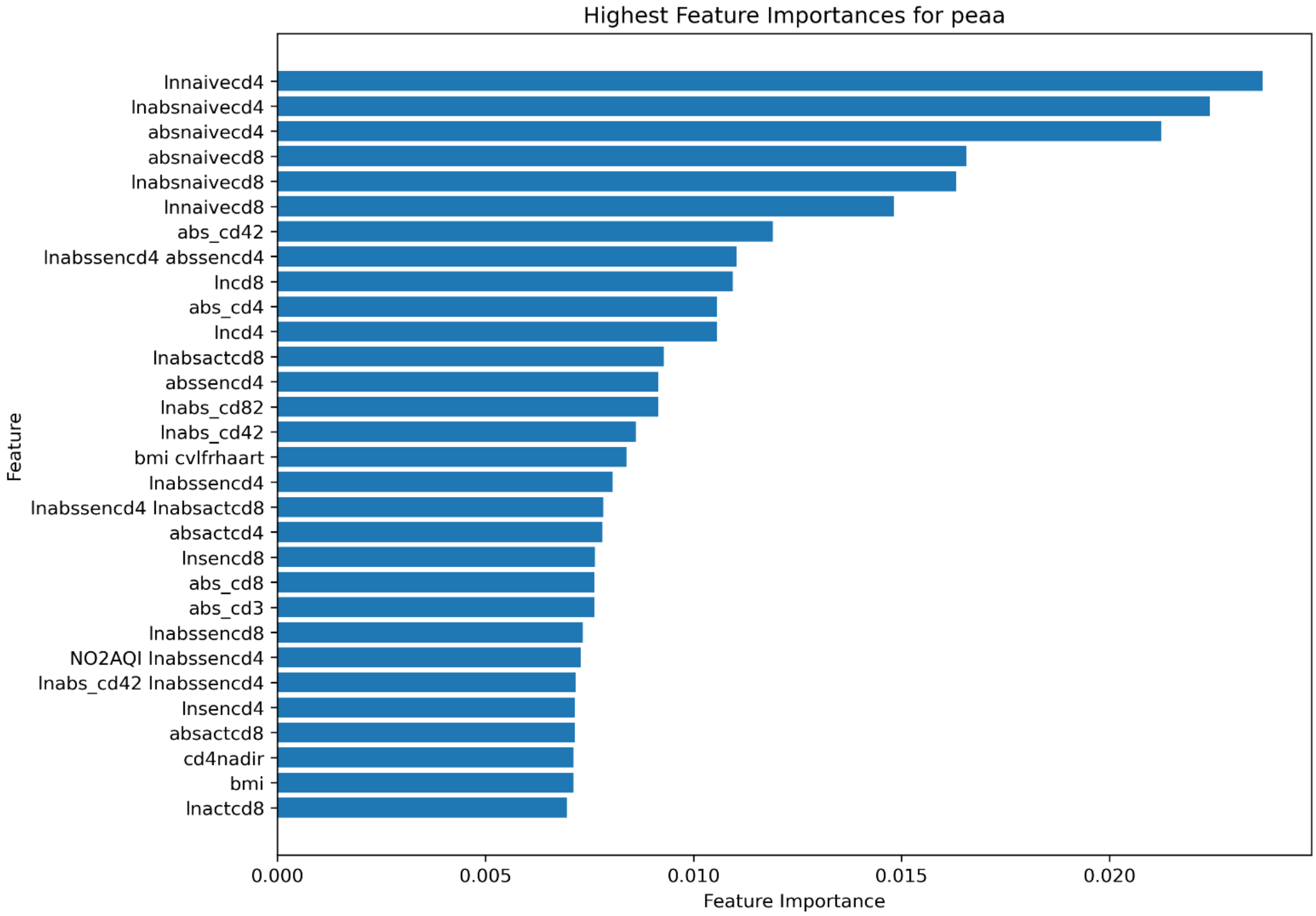

Highest Feature Importances for geaa
Feature
cum_pkyear
educbas
lnactcd4
white
lncd3
abssencd8
lnnaivecd4
absactcd8
lnactcd8
lnsencd8
lnabsactcd8
cd4nadir
bmi
lnabsnaivecd4
lncd8
absnaivecd4
SBMPM2.5
lncd4
lnsencd4
absactcd4
OzoneAQI
abs_cd3
lnabssencd4
lnabs_cd82
lnabssencd8
absnaivecd8
lnnaivecd8
SPWPM2.5
lnabsactcd4
abssencd4
0.00
0.01
0.02
0.03
0.04
0.05
0.06
0.07
0.08
Feature Importance



Highest Feature Importances for eeaa
Feature
absactcd8
lnabsactcd8
lnnaivecd8
lnnaivecd4
lncd8
lnabsnaivecd4
absnaivecd8
lncd4
lnactcd8
cvl
absnaivecd4
abssencd4
lnabsnaivecd8
lnabssencd8
abssencd8
cd4nadir lnabssencd4
abs_cd8
lnsencd8
NO2 lnabssencd4
lnabssencd4
abs_cd4
lnsencd4
lnabs_cd82
lnabs_cd42
SO2 lnabssencd4
SO2 abssencd8
educbas lnabssencd4
lnabssencd4 abssencd4
abs_cd3
SO2AQI educbas
0.000
0.005
0.010
0.015
0.020
0.025
Feature Importance



Highest Feature Importances for dnamtladjage
Feature
absactcd8
lnabsactcd8
lncd8
lnactcd8
lncd4
lnabssencd8
abssencd8
lnnaivecd8
lnabsnaivecd8
lnnaivecd4
absnaivecd4
lnabs_cd82
cvl
lnabsnaivecd4
lnabssencd4 abssencd4
absnaivecd8
lnsencd8
lnabssencd4
abs_cd8
abssencd4
lnsencd4
lnabssencd4 cvl
SO2 lnabssencd4
lncd3
lnabs_cd42
lnabssencd4 bmi
lnabssencd4 vload2
abs_cd3
abs_cd42
lnabssencd4 lnabssencd8
0.000
0.005
0.010
0.015
0.020
0.025
Feature Importance


Highest Feature Importances for peaa
Feature
lnnaivecd4
lnabsnaivecd4
absnaivecd4
absnaivecd8
lnabsnaivecd8
lnnaivecd8
abs_cd42
lnabssencd4 abssencd4
lncd8
abs_cd4
lncd4
lnabsactcd8
abssencd4
lnabs_cd82
lnabs_cd42
bmi cvlfrhaart
lnabssencd4
lnabssencd4 lnabsactcd8
absactcd4
lnsencd8
abs_cd8
abs_cd3
lnabssencd8
NO2AQI lnabssencd4
lnabs_cd42 lnabssencd4
lnsencd4
absactcd8
cd4nadir
bmi
lnactcd8
0.000
0.005
0.010
0.015
0.020
Feature Importance

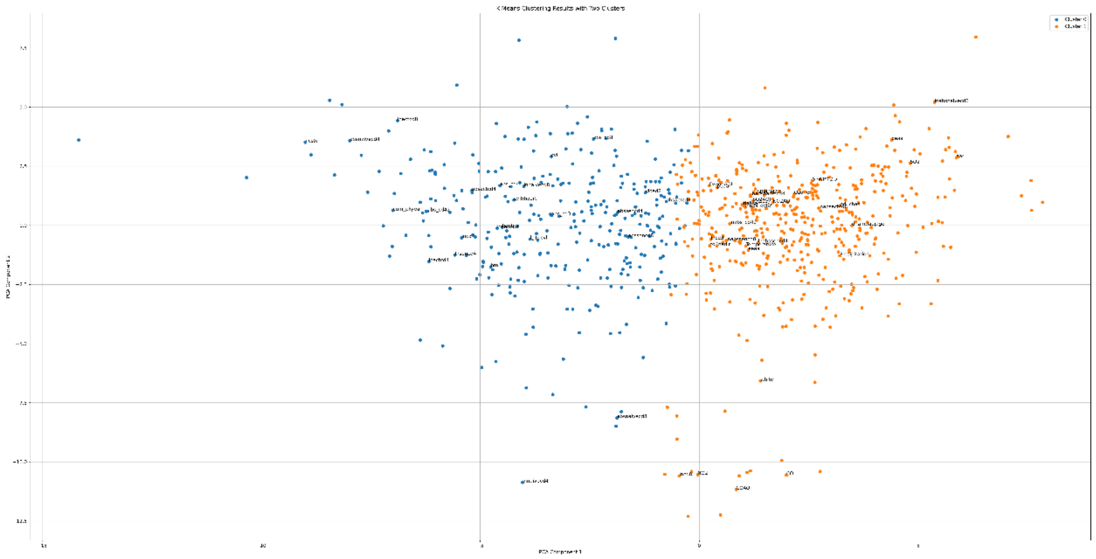

K-Means Clustering Results with Two Clusters
Cluster 0
Cluster 1
PCA Component 2
PCA Component 1
-15
-10
-5
0
5

Figure 3: K-Means Clustering Results please see attached files

Figure 4: DoWhy Results- Please see attached files

Figure 4: XGboost Results- Please see attached files

4. Discussion

In this study, we have used four different epigenetic clock measures of aging based on methylation patterns of genomic DNA, each of which calculates years of biological age acceleration relative to chronologic age. Additionally, we use an age-adjusted DNA methylation-based estimate of the telomere length at the ends of chromosomes. Incorporating climate and air quality data in addition to preexisting data on people living with and without HIV, by applying feature selection and machine learning methods, we demonstrate that epigenetic predictors are significantly associated with these measures of aging, while external factors like air quality and BMI have more marginal effects.

Epigenetic features have consistently shown high feature importance across all measures of epigenetic aging during feature selection. Increases in active and senescent CD4 and CD8 cell counts are generally associated with positive correlation on epigenetic aging, while increases in naive CD4 and CD8 cells are associated with negative effects on epigenetic aging. This makes sense, as the body strives to maintain CD3+ T-cell homeostasis over time. Therefore, as we age, decreases in naïve T-cell counts are associated with increases in active and senescent cell counts.

External factors appear to have less feature importance across all measures of epigenetic aging during feature selection. Except for GEAA, which has significantly more external factors ranking with high importance during the random forest, all other measures of aging have a select few external factors that appear with high importance except when doing K-means clustering. Smoking consistently appears as an extremely influential feature on GEAA, as it not only ranks first in feature importance, but also has increasing effects on GEAA when smoking increases. However, it is important to note that GEAA was built with smoking in mind, so it is not surprising that smoking is a significant feature when it comes to GEAA. To better understand the negative findings related to temperature and precipitation, it is crucial to clarify how these variables were evaluated and measured in the context of aging. Temperature and precipitation data were sourced from the National Centers for Environmental Information (NCEI) and were based on monthly averages. Temperature was measured using a scaling factor of 10 for precision, with the actual values converted to Celsius. Precipitation was similarly scaled, with values converted from millimeters to the actual precipitation levels by dividing by 10. These environmental factors were evaluated about different measures of epigenetic aging—including AAR, EEAA, PEAA, GEAA, and others. However, despite being considered, temperature and precipitation did not show significant correlations with any of the aging measures in this study. This lack of significant findings suggests that, over the study period and within the conditions of the analysis, these climate factors did not exert a notable influence on aging processes in the individuals evaluated. In contrast, air quality—specifically ozone concentration—was found to have a significant impact. The study showed a slight positive correlation between higher ozone concentrations and accelerated aging, as indicated by its effects on AAR and GEAA. This suggests that air pollution, particularly the presence of ozone, could have a more direct effect on aging, possibly due to its inflammatory and oxidative stress-inducing properties, which are known to influence epigenetic aging processes.

These findings highlight the complex nature of environmental influences on aging, with air quality appearing to be more impactful than temperature and precipitation in this analysis. This contrast underscores the importance of considering various environmental exposures and their potential differential effects on aging mechanisms.

This finding underscores the importance of regulating air pollution, as prolonged exposure to pollutants like ozone may accelerate epigenetic aging. Given the growing concerns about climate change, which could worsen air quality, policies to reduce air pollution, especially in urban areas and regions with high industrial activity, are crucial. Additionally, further research is needed to explore the biological mechanisms linking air quality and aging, and longitudinal studies should assess the long-term effects of pollutants. This study highlights the need for integrated approaches in public health and aging research, emphasizing environmental factors like air quality as key considerations for improving health outcomes and protecting vulnerable populations from accelerated aging.

The length of the MACS study, the involvement of people both infected and not infected with HIV, and its extensive epigenetic information on its participants makes it possible for us to examine the long-term effects of living and aging with HIV. Pairing with available climate and air quality data, we can utilize the MACS study to examine long-term relationships between HIV aging and these external factors. Future work will entail employing more advanced feature selection and causal inference models and formulating policy recommendations.

The MACS cohort used in this study only includes men who have sex with men. This limits our ability to generalize our results to women living with HIV. The MACS also only enrolled small numbers of non-white participants, meaning that it lacked sufficient statistical power to examine both initial HIV infection and race. Future enrollments for the MACS are intended to include a more diverse set of patients for future studies.

Data and code availability statement

Epigenetic data from the Multicenter AIDS Cohort Study (MACS) will be acquired upon approval of the concept sheet. Air quality and climate data have been sourced 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. The code is available here: https://github.com/ryanhu00/Epigenetic-Aging-with-External-Factors

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).

CC,RH,ECB,RS,MSE,FR,MM,JM,TB,SG,SDW,ECB,RMA,BDJ, and CMR gratefully acknowledge the contributions of the study participants and the dedication of the staff at the MWCCS sites.

RH, CC, and RMA acknowledge funding support from the Caltech Linde Center for Science, Society, and Public Policy.

B.D. Jamieson is supported by U01-HL146333 and M.E. Sehl was awarded the Susan G. Komen Career Catalyst Award CCR16380478.

C. Ramirez receives support from P30 MH058107.

Author Contribution Statement

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

CC developed the research question hypothesis, research design, and methodology. BDJ and CR edited it and supported data acquisition and HIV pathogenesis knowledge.

RH performed data analysis and interpreted the results. CC conducted the manuscript writing.

ECB performed study participant selection for the original study.

RMA, ECB, BDJ, and CR reviewed and edited it. RMA secured funding for the project, CR provided overall Supervision.

MS provided the original epigenetic clock data and interpretation of that data.

RS acquired the original epigenetic data and generated the database and calculation of epigenetic age for this analysis.

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.

Conflicts of interest

The authors declare no conflict of interest.

5. References

[1] Breen, E. C. et al. Accelerated aging with HIV occurs at the time of initial HIV infection. iScience 25, 104488 (2022). URL https://www.sciencedirect.com/science/article/pii/S2589004222007593.

[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 (2024). URL https://www.frontiersin.org/journals/bioinformatics/articles/10.3389/fbinf.2024.1357889.

[3] Merzouki, A., Estill, J., Orel, E., Tal, K. & Keiser, O. Clusters of sub-Saharan African countries based on sociobehavioral characteristics and associated HIV incidence. bioRxiv (2020). URL https://www.biorxiv.org/content/early/2020/12/18/620450.

[4] Mutai, C., McSharry, P., Innocent, N. & Musabanganji, E. Use of unsupervised machine learning to characterize HIV predictors in sub-Saharan Africa. BMC Infectious Diseases 23 (2023).

[5] 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).

[6] 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. PMID: 19537950,

[7] Bhavan, K. P., Kampalath, V. N. & Overton, E. T. The aging of the HIV epidemic. Current HIV/AIDS Reports 5, 150–158 (2008).

[8] Shiau, S. et al. Epigenetic aging biomarkers associated with cognitive impairment in older African American adults with human immunodeficiency virus (HIV). Clinical Infectious Diseases 73, 1982–1991 (2021).

[9] 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).

[10] Saldana, C. et al. Development of a machine learning modeling 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).

[11] Nkiruka, O., Prasad, R. & Clement, O. Prediction of malaria incidence using climate variability and machine learning. Informatics in Medicine Unlocked 22, 100508 (2021). URL https://www.sciencedirect.com/science/article/pii/S2352914820306596.

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

[13] 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).

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

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

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

Supplementary information (SI)

Appendix A: Feature selection process

For random forest, a Python script was written to select the top 30 features of the dataset related to each of the target variables. Applying the sci-kit-learn library, GridSearchCV was used for hyperparameter optimization to find the best set of parameters for performance. With the best parameters found for each target variable, a random forest was implemented to find the top 30 features, and the results were plotted for each target variable.

For K-means clustering, a slightly different process was used. Instead of running K-means clustering five times for each measure of epigenetic aging as was done with random forest, K-means clustering was implemented only once on the original set of features. This approach was taken because, unlike random forest, K-means clustering is an unsupervised method. As such, GridSearchCV was not a viable method for hyperparameter optimization. Instead, a different approach was employed. To determine the optimal number of clusters, a silhouette score was used. With the optimal number of clusters and best parameters identified, K-means clustering was performed.