**Final Research Paper**

Sherri Ricardes

Colorado State University Global

Course Code: MIS581

Dr. Steve Chung

May 4, 2025

Table of Contents

[ABSTRACT 4](#_Toc197283593)

[INTRODUCTION 6](#_Toc197283594)

[OBJECTIVES 6](#_Toc197283595)

[OVERVIEW OF STUDY 8](#_Toc197283596)

[RESEARCH QUESTIONS AND HYPOTHESES 9](#_Toc197283597)

[Research Question 1 9](#_Toc197283598)

[Research Question 2 9](#_Toc197283599)

[Research Question 3 10](#_Toc197283600)

[Overview of the Hypotheses 10](#_Toc197283601)

[LITERATURE REVIEW 11](#_Toc197283602)

[RESEARCH DESIGN 12](#_Toc197283603)

[Methodology 12](#_Toc197283604)

[Methods 13](#_Toc197283605)

[Limitations 14](#_Toc197283606)

[Ethical Considerations 14](#_Toc197283607)

[FINDINGS 14](#_Toc197283608)

[Research Question 1: 18](#_Toc197283609)

[Research Question 2: 19](#_Toc197283610)

[Research Question 3: 20](#_Toc197283611)

[Model Diagnostics 22](#_Toc197283612)

[Model Prediction 23](#_Toc197283613)

[Hypothesis Testing Summary 26](#_Toc197283614)

[CONCLUSION 27](#_Toc197283615)

[RECOMMENDATIONS 29](#_Toc197283616)

[REFERENCES 31](#_Toc197283617)

# ABSTRACT

As the global population continues to age, understanding the biological processes underlying aging has become increasingly important. This study aims to estimate biological age using a combination of molecular biomarkers, physiological indicators, and lifestyle factors drawn from the 2001–2002 National Health and Nutrition Examination Survey (NHANES). Using DNA methylation-based aging clocks (e.g., HorvathAge, GrimAge), telomere length, inflammation markers, metabolic health data, and self-reported health behaviors, the project applies both multiple linear regression and machine learning techniques to develop predictive models of biological age. In this study, GrimAge, a DNA methylation-based epigenetic clock, is used as the primary proxy for biological age due to its strong correlation with aging-related outcomes.

Separate models were constructed to address three research questions. The initial model using biomarkers alone was statistically significant but explained a modest proportion of variance (adjusted R² = 0.078). The addition of lifestyle factors, such as physical activity level, slightly improved model performance (adjusted R² = 0.089). The final comprehensive model, incorporating biomarkers, demographic factors, and self-reported health status, was highly significant, explaining 76.7% of the variance in GrimAge (adjusted R² = 0.764). Significant predictors included chronological age, telomere length, vitamin D levels, gender, CRP levels, and self-rated health. Higher chronological age, elevated CRP levels, and male gender were associated with increased biological age, while longer telomeres, higher vitamin D, and better self-rated health were linked to lower biological age estimates.

To improve prediction accuracy, machine learning models were evaluated. The Random Forest model performed very well, with a high R² score of 0.964 and a low RMSE of 1.727, indicating a strong fit to the data. The Bagged Trees model produced nearly identical results, demonstrating consistency across tree-based approaches. In both models, the most important predictors of biological age were chronological age, telomere length, BMI, weight, vitamin D, and gender. These results show that machine learning techniques are effective at capturing complex patterns in aging-related data.

Findings from this research highlight the complex and multi-dimensional nature of aging and suggest that modifiable factors, such as inflammation and lifestyle behaviors, may influence biological aging. These insights could guide future strategies for early intervention, personalized healthcare, and public health programs aimed at promoting healthy aging.

# INTRODUCTION

As populations worldwide continue to live longer, understanding the biological drivers of aging has become very important. Healthy aging is not just about extending lifespan but improving quality of life as people grow older. As we age, our bodies experience changes at the cellular and molecular levels, and some individuals experience accelerated aging due to factors like chronic stress, inflammation, hormone imbalances, and lifestyle choices.

Recent advances in data analytics and machine learning have opened new opportunities for predicting health outcomes based on biological markers, or biomarkers. Biomarkers such as inflammation markers, blood sugar levels, and other physiological indicators could provide insight into a person’s biological age, which may differ from their chronological age. Understanding these factors could help healthcare professionals and individuals create strategic health plans that target specific areas of concern and slow down aging.

This research area is important because it has the potential to shift how we approach aging, from reactive treatments to proactive prevention. If we can predict biological age and understand the key drivers of accelerated aging, it could help reduce the risk of age-related diseases like heart disease, diabetes, and cognitive decline. Ultimately, this contributes to improving public health, reducing healthcare costs, and supporting wellness programs aimed at healthy aging.

# OBJECTIVES

The primary goal of thisresearch is to develop a predictive modelusing linear regression and machine learningthat estimates biological age based on biomarker and lifestyle data. This project uses data from the 2001–2002 National Health and Nutrition Examination Survey (NHANES), which isa nationally representative dataset that includes clinical, demographic, and laboratory information.

Specific objectives include:

* Identifying key biomarkers and lifestyle factors associated with biological aging.
* Examining how health and lifestyle factors are related to biological age, based on estimates from different epigenetic clocks (e.g., HorvathAge, GrimAge).
* Comparing the performance of statisticalmethods, linear regression, and machine learning models in estimating biological age.
* Providing insights that support health strategies aimed at slowing the aging process and improving wellness.

The key variables analyzed from the NHANES dataset are shown in Table 1.

**Table 1**

*Key variables analyzed from the NHANES dataset.*

|  |  |
| --- | --- |
| **Variable** | **Description** |
| Sequence Number | Unique identifier |
| Telomere Length | A classical biomarker of cellular aging |
| DNA Methylation | Including GrimAge, derived from methylation datasets |
| C-reactive Protein (CRP) | A marker of systemic inflammation |
| Glucose and Insulin | Indicators of metabolic health and insulin sensitivity |
| Cholesterol Markers | LDL and triglycerides, linked to cardiovascular risk |
| Vitamin D | Important for immune regulation and bone he |
| BMI and Weight | Reflects body composition and weight |
| Physical Activity and Fitness Level | Modifiable factors linked to healthy aging |
| General Health Status | Self-reported measure of well-being |
| Demographics | Includes age, gender, and education level |

By integrating these variables into predictive models, the project aims to better understand how biology and lifestyle interact in the aging process and to help shape future health interventions that are more personalized and proactive.

# OVERVIEW OF STUDY

This study aims to explore the relationship between biological age and a range of health and lifestyle variables using data from the 2001–2002 cycle of the National Health and Nutrition Examination Survey (NHANES). NHANES is a large, publicly available dataset that combines laboratory results, physical examinations, and self-reported questionnaires to assess the health and nutritional status of adults and children in the United States. For this project, multiple NHANES datasets were merged using a common participant ID (SEQN) to create a single, comprehensive dataset for analysis.

The final merged dataset includes biological markers such as telomere length, inflammatory markers (e.g., C-reactive protein), glucose and insulin levels, cholesterol levels (LDL and triglycerides), and vitamin D. It also incorporates physical health indicators such as Body Mass Index (BMI), weight, and cardiovascular fitness level. Lifestyle and demographic variables such as physical activity, self-rated health status, education level, age, and gender are also included to provide a well-rounded picture of the individual.

The primary objective is to develop and evaluate predictive models of biological age using both linear regression and machine learning algorithms. Biological age is estimated using multiple DNA methylation-based aging clocks included in the NHANES methylation dataset, such as HorvathAge, GrimAge, and PhenoAge. These epigenetic clocks are increasingly used in research to provide a more accurate reflection of physiological aging than chronological age alone. In this study, GrimAge serves as the primary proxy for biological age. Prior research has demonstrated that GrimAge is strongly associated with morbidity and mortality risk, making it a robust and clinically relevant measure for capturing age-related biological decline.

Exploratory data analysis is conducted to assess variable distributions, identify outliers, and examine correlations among variables. Linear regression is applied to evaluate how individual biomarkers and lifestyle factors relate to biological age. Also, machine learning models are used to test whether combining multiple variables can better estimate biological age, with performance measured using metrics such as R-squared and cross-validation scores.

This approach allows for an in-depth study of the biological aging process and supports the identification of key modifiable factors associated with accelerated aging. The insights gained can help guide health strategies, wellness programs, and public health initiatives aimed at promoting healthy aging.

# RESEARCH QUESTIONS AND HYPOTHESES

## Research Question 1

**Can specific biomarkers (e.g., telomere length, DNA methylation, cholesterol, glucose, insulin) be used to estimate a person’s biological age?**

* Null Hypothesis (H0): There is no linear relationship between the selected biomarkers (e.g., telomere length, DNA methylation, cholesterol, glucose, insulin) to estimate biological age.
* Alternate Hypothesis (H1): There is a linear relationship between the selected biomarkers (e.g., telomere length, DNA methylation, cholesterol, glucose, insulin) to estimate biological age.

## ****Research Question 2****

**Do lifestyle factors, such as physical activity and fitness level, add predictive value to estimated biological age beyond what is explained by biomarkers alone?**

* Null Hypothesis (H0): There is no relationship between lifestyle factors (physical activity level, fitness level) and estimated biological age beyond what is explained by biomarkers.
* Alternate Hypothesis (H1): There is a relationship between lifestyle factors (physical activity level, fitness level) and estimated biological age beyond what is explained by biomarkers.

## ****Research Question 3****

**Is the mean difference between biological age and chronological age significantly greater in individuals with high inflammation, high cholesterol, poor glucose/insulin control, and low physical activity compared to those with healthier profiles?**

* Null Hypothesis (H0): There is no difference in the **mean age**(biological age minus chronological age) between individuals with poor health indicators (e.g., high CRP, high cholesterol, poor glucose/insulin control, low physical activity) and those with healthier profiles.
* Alternate Hypothesis (H1): There is a difference in the **mean age**(biological age minus chronological age) between individuals with poor health indicators (e.g., high CRP, high cholesterol, poor glucose/insulin control, low physical activity) and those with healthier profiles.

## ****Overview of the Hypotheses****

These hypotheses are based on research that shows certain biomarkers can help estimate how someone is aging. For example, DNA methylation-based measures, such as HorvathAge and GrimAge, have emerged as some of the most accurate indicators of biological age, reflecting molecular changes that occur with aging. Telomere length is another well-known biomarker, and other factors like inflammation (e.g., CRP), cholesterol, and blood sugar control also play important roles in long-term health. When you add in lifestyle factors like physical activity or body weight, you begin to see a more complete picture of the aging process. This project’s goal is to explore whether these combined factors, such as molecular, physiological, and behavioral, are associated with estimated biological age and to examine how they work together to explain differences in how people age.

# LITERATURE REVIEW

When we think about aging, we usually think of how many birthdays someone has had, which is chronological age. But that number doesn’t always reflect how someone is doing on the inside. Some people stay healthier and more energetic well into old age, while others experience age-related issues much earlier. That’s where the concept of biological age comes in. Biological age aims to measure how fast a person is actually aging based on physical, cellular, and molecular signs in the body. Recent research shows that combining different health markers with advanced analytical methods, such as linear regression and machine learning, can provide a more accurate and personalized picture of how a person is truly aging. These models help identify patterns in biomarker and lifestyle data that may not be obvious on the surface, offering deeper insight into an individual's biological age beyond what chronological age alone can reveal.

One of the most studied markers of biological aging is telomere length, which shortens with each cell division and is influenced by stress, inflammation, and environmental exposures. Vaiserman and Krasnienkov (2021) emphasize that while telomere length alone has limitations, it remains a valuable biomarker when combined with others like DNA methylation. Also, Ni et al. (2023) found that predictive models incorporating telomere length alongside additional health data, such as cholesterol and glucose levels, improved accuracy in estimating biological age.

Researchers agree that no single marker is sufficient. Erema et al. (2022) and Bortz et al. (2023) both stress the importance of combining clinical, molecular, and epigenetic data for the most accurate predictions. Bortz et al. used UK Biobank data and showed that machine learning models built on routine lab values (e.g., Vitamin D, Cystatin C, ALT) outperformed older models like PhenoAge. This aligns with Ni et al.'s conclusions that models using common health data and modern machine learning techniques, such as Bagged Trees and Rational Quadratic regression, can produce strong results even in different demographic groups.

Jia et al. (2017) explored several estimation methods, such as multiple linear regression and principal component analysis, and highlighted that approaches like the Klemera and Doubal Method (KDM) tend to be more accurate because they incorporate information from various physiological systems. Their findings support the idea that using a wide range of data, from immune response to metabolic and inflammatory markers, can lead to more reliable predictions of biological age.

This research supports the main idea behind my project: that the best way to estimate someone’s biological age is by combining different health markers and using methods like linear regression, machine learning, and statistical modeling. Taking this more complete approach gives a better understanding of how someone is aging and can also help healthcare providers catch potential problems early and offer more personalized care.

# RESEARCH DESIGN

## Methodology

This project uses a quantitative research design to estimate biological age by analyzing a combination of molecular, physiological, and lifestyle variables. The primary data source is the NHANES 2001–2002 dataset, which includes key indicators such as telomere length, inflammatory and metabolic biomarkers (e.g., CRP, glucose, insulin, cholesterol), fitness and physical activity levels, general health status, weight, and demographics.

In addition to these traditional biomarkers, the analysis incorporates DNA methylation-based aging clocks, specifically HorvathAge, HannumAge, PhenoAge, and GrimAge. These epigenetic clocks are well-established in the scientific literature for their ability to quantify biological aging based on methylation patterns across the genome.

The primary analytical method is multiple linear regression, selected for its interpretability and ability to assess the strength and direction of relationships between predictors and the outcome variable. Linear regression also provides a baseline for comparing model performance to more advanced machine learning methods. To explore potential non-linear relationships and enhance prediction accuracy, machine learning models, including Bagged Trees, Random Forests, and Support Vector Machines (SVMs), were also evaluated in this study. While these models demonstrated higher predictive accuracy, they are generally less interpretable than linear regression. Model performance was evaluated using R-squared, root mean square error (RMSE), and mean absolute error (MAE).

## Methods

Data preparation began with handling missing values through mean imputation to maintain consistency across variables. Continuous variables were normalized where necessary, and categorical variables such as gender, general health status, and fitness level were encoded using dummy variables. Feature selection techniques, including correlation analysis and variance inflation factor (VIF) checks, were used to identify relevant predictors while reducing multicollinearity.

The dataset was analyzed without splitting into separate training and testing sets, as the primary objective was explanatory modeling rather than predictive generalization. All analyses were performed in Python, using Statsmodels to run multiple linear regression and evaluate model performance through diagnostic checks. In addition to linear regression, machine learning algorithms, such as Random Forest, Bagged Trees, and Support Vector Machines, were also applied to explore non-linear relationships and improve prediction accuracy. These models were evaluated using R², RMSE, MAE, and feature importance metrics to compare their performance against the regression baseline.

## Limitations

The NHANES dataset provides a diverse set of health-related variables, making it a strong foundation for research on aging. However, there are some limitations. First, not all biomarkers or aging clock measures are available for every participant, which can lead to missing data and a smaller usable sample size after preprocessing. Second, NHANES shows a single point in time for each person, making it harder to understand how aging changes over the years. Also, self-reported information, such as physical activity levels or general health status, may be prone to bias or inaccuracies. Finally, while statistical and machine learning methods are useful for finding patterns, they might not fully capture the complex and changing ways that different biological systems interact as we age.

## Ethical Considerations

Even though NHANES data are publicly accessible and de-identified, working with sensitive health information still needs thoughtful ethical consideration. If this study were based on original data collection, informed consent would be essential, along with clear communication about how participants’ health and biological data would be used. Ensuring privacy and securing the data are critical, as well as transparency around how predictive models might be applied. Also, it's important not to exaggerate how accurate or meaningful biological age estimates are, especially if they might affect decisions about healthcare, jobs, or insurance.

# FINDINGS

Before conducting the regression analyses, descriptive statistics were generated for all numeric variables included in the dataset. The dataset contained 6,621 observations and 21 variables after merging multiple NHANES data sources and preprocessing. Summary statistics, including the mean, standard deviation, minimum, maximum, and quartiles, were calculated for key variables such as BMI, triglycerides, LDL, C-reactive protein (CRP), telomere length, glucose, insulin, vitamin D, weight, age, GrimAge (biological age estimate), and others. See Figure 1. These descriptive statistics provided an overview of the sample’s health status and helped identify missing data patterns and variable distributions prior to modeling.

**Figure 1**

*Summary Statistics of Numeric Variables in the Dataset.*

*A screenshot of a computer screen

Description automatically generated*

The correlation matrix in Figure 2 provides an overview of the relationships among numeric variables in the dataset. As expected, several DNA methylation-based biological age measures (e.g., HorvathAge, HannumAge, PhenoAge, GrimAgeMort) showed strong positive correlations with each other (r > 0.75), reflecting their common purpose of estimating biological age. Chronological age was moderately to strongly correlated with several biological age indicators (e.g., r = 0.83 with HorvathAge, r = 0.77 with GrimAgeMort), supporting the validity of these measures. Telomere length was weakly and negatively correlated with age (r = –0.43), consistent with the biological expectation that telomeres shorten over time. Other health biomarkers, such as glucose, insulin, and C-reactive protein (CRP), showed weaker correlations with biological age measures but still contribute important independent information for modeling. Overall, the matrix highlighted both expected biological patterns and the largely independent contribution of lifestyle and metabolic variables.

**Figure 2**

*Correlation matrix of Numeric Variables in the Final Dataset.*

A screenshot of a graph

Description automatically generated

The pair plot in Figure 3 visualizes the relationships between selected key variables, which includes chronological age, telomere length, CRP, vitamin D, and GrimAgeMort (estimated biological age). As expected, chronological age and GrimAgeMort displayed a strong positive association, while telomere length showed a negative relationship with both age and GrimAgeMort, consistent with known aging biology. CRP and GrimAgeMort were moderately positively associated, suggesting that higher inflammation levels may contribute to accelerated biological aging. Vitamin D appeared weakly negatively related to GrimAgeMort, indicating a possible protective role. The distributions along the diagonal further confirmed that most variables were approximately normally distributed, although CRP showed a right-skewed distribution.

**Figure 3**

*Pair plot of Key Variables Related to Biological Age.*

A graph of a diagram

Description automatically generated with medium confidence

To address the research questions, three linear regression models were developed. The first model examined whether specific biomarkers, such as telomere length, cholesterol, and glucose, could be used to predict biological age. The second model added physical activity level to determine whether lifestyle factors improved the prediction. The third and final model incorporated a broader set of health indicators, including inflammation markers, vitamin D levels, weight, and self-reported health status, to assess the relationship between overall health and biological aging. In addition to these regression models, a fourth model using machine learning (e.g., Random Forest) was also evaluated to test whether non-linear techniques could improve prediction accuracy. Each model’s results are discussed in the following sections, along with conclusions regarding the acceptance or rejection of the null hypotheses.

## Research Question 1:

Can specific biomarkers (e.g., telomere length, cholesterol, glucose, insulin) be used to estimate biological age?

* A multiple linear regression was conducted using biomarkers only: telomere length, triglycerides, LDL, glucose, and insulin.
* The model was statistically significant, F(5, 1302) = 23.02, p < 0.001.
* However, the model explained a small proportion of the variance in GrimAge (adjusted R² = 0.078).
* Telomere length was a statistically significant predictor (p < 0.001), with longer telomeres associated with lower biological age.
* Other predictors (triglycerides, LDL, glucose, insulin) were not statistically significant (p > 0.05).

Based on these results, the null hypothesis was rejected for telomere length, but failed to reject for triglycerides, LDL, glucose, and insulin. See Figure 4 for Research Question 1 results.

**Figure 4**

*Regression results for Research Question 1: Biomarkers only.*

*A screenshot of a computer

Description automatically generated*

## Research Question 2:

Do lifestyle factors, such as physical activity, add predictive value beyond biomarkers alone?

* Physical activity level was added to the biomarker model.
* The expanded model was statistically significant, F(6, 1301) = 22.33, p < 0.001, with a slightly higher adjusted R² = 0.089.
* Physical activity level was a statistically significant predictor (p < 0.001), with lower activity levels associated with higher biological age estimates.

Based on these results, the null hypothesis was rejected for the addition of physical activity. See Figure 5 for Research Question 2 results.

**Figure 5**

*Regression results for Research Question 2: Biomarkers plus lifestyle factors.*

*A screenshot of a computer

Description automatically generated*

## Research Question 3:

Is the mean difference between biological and chronological age greater among individuals with poor health indicators?

* A full model was developed including inflammation (CRP), vitamin D, weight, BMI, age, gender, and self-reported health.
* The final model was statistically significant, F(17, 1290) = 249.5, p < 0.001, explaining 76.7% of the variance in GrimAge (adjusted R² = 0.764).
* Significant predictors included chronological age, telomere length, CRP, vitamin D, gender, and self-reported health status (excellent/very good).
* Higher chronological age, elevated CRP levels, and male gender were associated with increased biological age, while longer telomere length, higher vitamin D levels, and better self-rated health were associated with decreased biological age.

Based on these results, the null hypothesis was rejected for CRP, vitamin D, and self-reported health. See Figure 6 for Research Question 3 results.

**Figure 6**

*Regression results for Research Question 3: Full model with biomarkers, lifestyle, and health indicators.*

*A screenshot of a computer

Description automatically generated*

In addition to traditional linear models, a Random Forest machine learning model was tested and showed substantially better predictive accuracy, with an R² of 0.9640, a Root Mean Squared Error (RMSE) of 1.7272, and a Mean Absolute Error (MAE) of 2.4567. See Figure 7. This suggests that non-linear modeling approaches may be better suited for capturing the complex biological and lifestyle factors that influence aging. These findings support the use of advanced analytics in aging research and highlight the importance of continued exploration of machine learning tools in future studies**.**

**Figure 7**

*Random Forest Model Results: R² Score, RMSE, MAE, and Top 10 Feature Importance.*

*A screenshot of a computer

Description automatically generated*

To evaluate which modeling approach most effectively estimates biological age, the performance of multiple predictive models was compared. Among these, the Bagged Trees and Random Forest models demonstrated the highest accuracy, producing significantly higher R² scores and lower RMSE values than both Linear Regression and Support Vector Machine models, as shown in Figure 8.

**Figure 8**

*Model performance comparison using R² (blue bars) and RMSE (red line). Bagged Trees and Random Forest outperformed Linear Regression and SVM in predicting biological age.*

**A graph with a red line

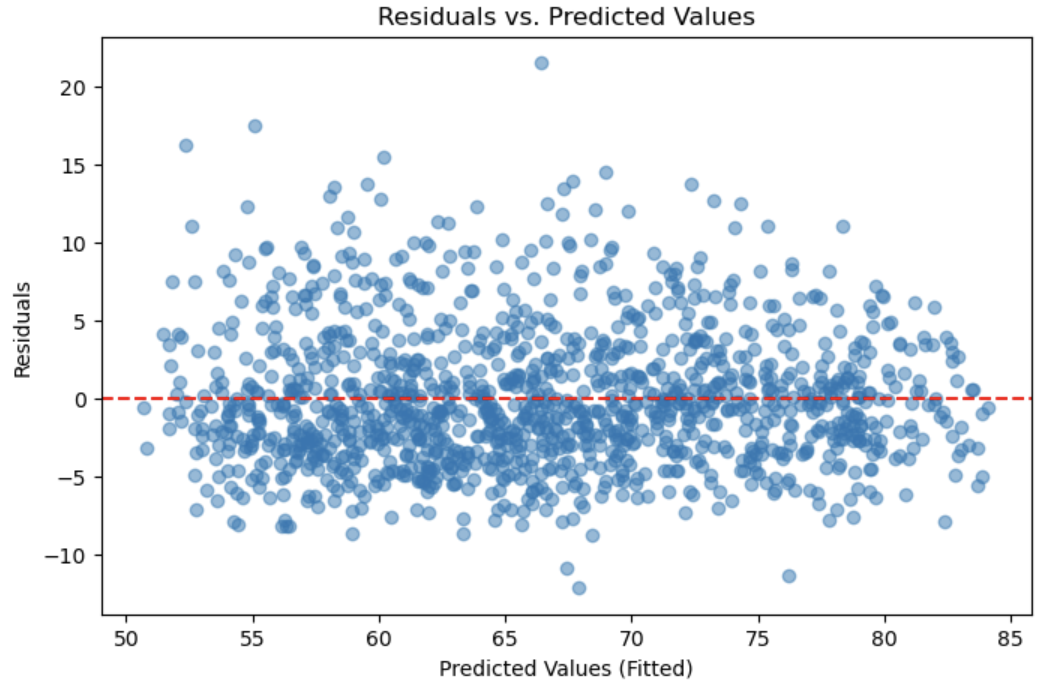
Description automatically generated**

## Model Diagnostics

To evaluate whether the assumptions of linear regression were reasonably met for the full model, residual diagnostics were performed. The residuals versus predicted values plot (See Figure 9) showed that residuals were centered around zero across all predicted values, suggesting that the model captured the data reasonably well without clear patterns of bias.

**Figure 9**

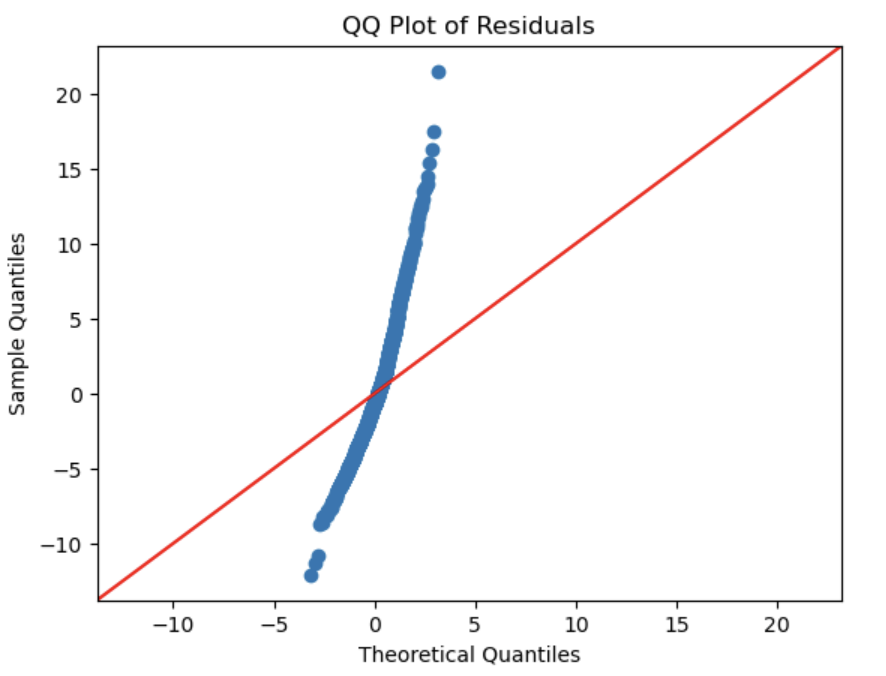
*Residuals vs. Predicted Values for the Full Model for Linear Regression.*



However, the QQ plot of residuals (see Figure 10) shows that the model assumption of normality wasn’t met. The residuals didn’t follow the expected straight-line pattern, especially at the high and low ends, which means there were some extreme values the model didn’t handle well. Because of this, the linear regression results may not fully reflect the true patterns in the data. This suggests that using other methods, such as machine learning models, could give better and more accurate predictions.

**Figure 10**

*QQ Plot of Residuals for the Full Model for Linear Regression.*



## Model Prediction

The formula for predicting biological age (GrimAge) using the full model is based on the regression coefficients. In general, the model prediction follows the structure:

**Predicted biological age** = (constant) + (coefficient₁ × variable₁) + (coefficient₂ × variable₂) + ... + (coefficientₙ × variableₙ)

For this study, the final equation for predicting biological age is:

Predicted biological age = 23.8916

− 1.6248 × TelomereLength

+ 0.0016 × Triglycerides

− 0.000004206 × LDL

− 0.0834 × Glucose

− 0.0393 × Insulin

− 0.2743 × PhysicalActivityLevel

+ 1.0811 × CReactiveProtein

− 0.1818 × VitaminD

− 0.0249 × Weight

+ 0.7396 × Age

+ 0.3795 × Gender\_numeric

− 2.2541 × GeneralHealth\_Excellent

− 1.1448 × GeneralHealth\_Fair

+ 0.3706 × GeneralHealth\_Good

+ 0.8029 × GeneralHealth\_Poor

− 1.2616 × GeneralHealth\_VeryGood

Each coefficient represents how much biological age is expected to change for a one-unit increase in the corresponding variable, assuming all other variables are held constant. Once the model is trained, new biological age predictions can be made using the same input variables as in the linear regression mode

Unlike linear regression, which provides a direct mathematical formula for prediction, the Random Forest model makes predictions using an ensemble of decision trees. Each tree in the forest is trained on a different subset of the data and captures patterns and interactions between predictors that may be non-linear or complex. To estimate biological age, the Random Forest model takes the average prediction from all trees in the ensemble, resulting in a robust and flexible prediction.

## Hypothesis Testing Summary

Across all three research questions, several null hypotheses were rejected based on the statistical results for linear regression:

* Research Question 1: The null hypothesis was rejected for telomere length, indicating that it is a significant predictor of biological age. However, the null hypothesis could not be rejected for triglycerides, LDL, glucose, and insulin, suggesting that these markers alone were not strong predictors in this dataset.
* Research Question 2: The addition of physical activity level led to a statistically significant improvement in the model, resulting in the rejection of the null hypothesis for lifestyle factors. This suggests that physical activity independently contributes to biological aging, beyond traditional biomarkers.
* Research Question 3: In the full model, the null hypothesis was rejected for CRP, vitamin D, and self-reported health status (excellent/very good). Chronological age and gender were also highly significant predictors. These findings support the idea that inflammation, vitamin D status, and perceived health play important roles in biological aging.

While some of the hypotheses were rejected, the models that used only biomarkers or biomarkers plus lifestyle factors explained only a small amount of the differences in biological age (adjusted R² = 0.078 and 0.089, respectively). This suggests that other important factors not included in this dataset, such as markers of oxidative stress, mitochondrial health, or genetic influences, might also play a big role in how people age. On the other hand, the full model, which combined a wider range of health indicators, explained much more of the variation in biological age (adjusted R² = 0.764). This highlights the importance of looking at multiple body systems together when trying to understand the aging process.

However, even this model had some limitations, such as non-normal residuals and possible non-linear relationships. To address these issues, machine learning models were also tested. Both the Random Forest and Bagged Trees models performed better, with R² scores of 0.964 and lower RMSE values, suggesting they captured patterns the linear models missed. These results also support the rejection of the null hypothesis across all three research questions when using machine learning, further validating the relationships between biological age and the selected health indicators. This reinforces the benefit of using more flexible methods when studying something as complex as biological aging.

It is important to note that a considerable amount of missing data required mean imputation for several variables, which may have reduced model precision. There were also other limitations that could have influenced the results, like possible errors in measurement, people not reporting their physical activity or health accurately, and the fact that the NHANES data captures just one point in time rather than following people over many years. Future studies should try to use long-term (longitudinal) data and include a wider range of biomarkers to build better models for predicting biological age and to better understand how aging really works.

# CONCLUSION

This project used a data-driven approach to estimate biological age by combining traditional health biomarkers with advanced measures such as DNA methylation-based aging clocks. By applying multiple linear regression models across different combinations of molecular, physiological, and lifestyle variables, the study uncovered important relationships that contribute to our understanding of the aging process. While the early models that focused only on biomarkers or a mix of biomarkers and lifestyle factors explained only a small part of the differences in biological age, the full model, which combined inflammation markers, vitamin D levels, self-reported health, demographics, and lifestyle factors, performed much better. These results highlight that biological aging is not driven by just one or two factors but is shaped by many different systems in the body working together, along with health behaviors that people can change.

In addition to traditional statistical models, the use of machine learning techniques, particularly Random Forest and Bagged Trees, further improved prediction accuracy and demonstrated their value in capturing more complex patterns in the data. While these models outperformed linear regression in terms of accuracy, their complexity can make them more difficult to interpret. However, these tools may be especially useful in developing future biological age assessments that are more personalized and precise.

Even though there were some challenges, such as missing data, possible measurement errors, and the fact that the NHANES dataset only captures one moment in time, the results still show strong evidence that certain factors, such as inflammation, telomere length, vitamin D, and physical activity, are important in how people age biologically. One of the biggest takeaways from this project is that it is not enough to look at just one health marker. Instead, the best way to understand aging is by building models that include many areas of health and lifestyle together.

Moving forward, this research lays important groundwork for creating more advanced models that could be used in healthcare and public health settings to better understand aging. Future studies should build on this by using more powerful machine learning techniques, working with long-term (longitudinal) data to track how aging changes over time, and expanding the types of biomarkers studied to include new measures like oxidative stress and mitochondrial function. By taking these steps, researchers can get closer to building practical tools that help catch early signs of accelerated aging, support healthier aging, and improve people’s quality of life as they grow older.

In summary, this project shows that understanding aging requires looking at the whole picture, how different parts of our body and lifestyle work together. It opens up new possibilities for combining data science, biology, and healthcare to help people age healthier.

# RECOMMENDATIONS

Based on the findings of this study, several recommendations can be made to enhance future research and practical applications related to biological aging:

1. Expand Modeling Techniques: While multiple linear regression provided useful insights, this study showed that machine learning models, such as Random Forest and Bagged Trees, offered significantly better prediction accuracy. Future research should build on this by exploring additional non-linear models, such as neural networks or gradient boosting methods, which may capture complex interactions between biomarkers and lifestyle factors more effectively than traditional approaches.
2. Follow Participants Over Time: Since NHANES provides only a snapshot of each participant’s health, future research should focus on datasets that follow individuals across their lifespan to directly observe how biological markers and health behaviors change with aging.
3. Expand the Range of Biomarkers: Future research should consider adding other health markers, such as those related to cell stress, energy production, or DNA repair. Including these newer types of biological signals could give a fuller picture of how aging happens inside the body and help improve how accurately we can predict biological age.
4. Consider Lifestyle and Psychosocial Variables: Future models could benefit from integrating more detailed lifestyle data, such as dietary patterns, sleep quality, stress levels, and socioeconomic factors, which have all been shown to influence biological aging.
5. Focus on Practical Applications: Future work should focus on turning biological age models into simple tools that doctors or wellness programs can use. Creating easy, non-invasive ways for people to check their biological age could help them track how they’re aging and make better lifestyle choices to stay healthier longer.
6. **Promote Healthy Lifestyle Programs:** Because factors like physical activity, healthy eating, and reducing inflammation were linked to slower biological aging, public health initiatives should continue to promote these behaviors to help support healthy aging across the broader population.

# REFERENCES

Berghoff, M. (n.d.). How can you measure biological age? Retrieved March 22, 2025,

from https://www.age.mpg.de/211116/how-can-you-measure-biological-age

Biomarkers of Aging Consortium, Herzog, C. M. S., Goeminne, L. J. E., et al. (2024).

Challenges and recommendations for the translation of biomarkers of aging. *Nature Aging*, 4(7), 1372–1383. https://doi.org/10.1038/s43587-024-00683-3

Bortz, J., Guariglia, A., Klaric, L., & et al. (2023). Biological age estimation using circulating

blood biomarkers. *Communications Biology*, 6, 1089. https://doi.org/10.1038/s42003-023-05456-z

Cui, H., Kong, Y., & Zhang, H. (2012). Oxidative stress, mitochondrial dysfunction, and aging.

Journal of signal transduction, 2012, 646354. https://doi.org/10.1155/2012/646354

Erema, V. V., Chervova, N. S., & Osipyants, A. I. (2022). Biological age predictors: The status

quo and future trends. *International Journal of Molecular Sciences*, 23(23), 15103. https://doi.org/10.3390/ijms232315103

Floridi, L., & Taddeo, M. (2016). What is data ethics? Philosophical Transactions of the Royal

Society A: Mathematical, Physical and Engineering Sciences, 374(2083), 20160360.

https://doi.org/10.1098/rsta.2016.0360

Jia, L., Zhang, W., & Chen, X. (2017). Common methods of biological age estimation. *Clinical*

*Interventions in Aging*, 12, 759–772. https://doi.org/10.2147/CIA.S134921

Jylhävä, J., Pedersen, N. L., & Hägg, S. (2017). Biological Age Predictors. EBioMedicine, 21,

29–36. https://doi.org/10.1016/j.ebiom.2017.03.046

National Health and Nutrition Examination Survey (NHANES) 2001-2002. Centers for Disease

Control and Prevention. Retrieved March 29, 2025, from https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2001

Ni, X., Zhao, H., Li, R., Su, H., Jiao, J., Yang, Z., Lv, Y., Pang, G., Sun, M., Hu, C., & Yuan, H.

(2023). Development of a model for the prediction of biological age. *Computer Methods and Programs in Biomedicine*, 240, 107686. https://doi.org/10.1016/j.cmpb.2023.107686

Polonsky, M. J., & Waller, D. S. (2019). Designing and managing a research project: A business

student's guide (4th ed.). SAGE Publications. ISBN: 9781544316468

Vaiserman, A., & Krasnienkov, D. (2021). Telomere Length as a Marker of Biological Age:

State-of-the-Art, Open Issues, and Future Perspectives. *Frontiers in*

*Genetics*. https://www.frontiersin.org/articles/10.3389/fgene.2020.630186/full