

2025



**Report prepared by**

Brandt Levitt

Allison E. Aiello

Audrey Kelly

Chantel L. Martin

Lauren Gaydosh

Kathleen Mullan Harris

# Wave V Epigenetic Clocks User Guide



CAROLINA POPULATION CENTER | CAROLINA SQUARE - SUITE 210 | 123 WEST FRANKLIN STREET | CHAPEL HILL, NC 27516

<https://doi.org/10.17615/05w9-7j15>

## Acknowledgments

The data from this study were generated as part of “The Add Health Epigenome Resource: Life course stressors and epigenomic modifications in adulthood” project funded by the National Institute on Minority Health and Health Disparities of the National Institutes of Health (R01 MD013349 to Kathleen Mullan Harris and Allison E. Aiello, MPIs). Data from Waves I-V of Add Health are from the Add Health Program Project, grant P01 HD31921 (Kathleen Mullan Harris) from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations. Add Health was originally designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill. Add Health is currently directed by Robert A. Hummer; it was previously directed by Kathleen Mullan Harris (2004-2021) and J. Richard Udry (1994-2004). Information on obtaining Add Health data is available on the project website (<https://addhealth.cpc.unc.edu>).

## Citation for User Guide

Levitt, B; Aiello, AE; Kelly, AL; Martin CL, Gaydos, L; Harris, KM. 2025. Wave V Epigenetic Clocks. Chapel Hill, NC: Carolina Population Center, University of North Carolina at Chapel Hill. Available from: <https://doi.org/10.17615/05w9-7j15>

## Table of Contents

1. Introduction.....	1
1.1 Overview of the Add Health Study .....	1
1.2 Rationale.....	1
1.3 Purpose of This Document .....	2
2. Background .....	3
2.1 Add Health and the Methylation Ancillary Study .....	3
2.2 Related Documentation.....	3
3. Overview of Variables.....	4
4. Data Processing Prior to Clock Calculation .....	5
4.1 Preprocessing of Beta Values .....	5
4.2 Quality Control Steps.....	5
4.3 Missing CpGs.....	5
5. Detailed Methods by Clock.....	7
5.1 Houseman Immune Cell Estimates.....	7
5.2 Horvath1 .....	7
5.3 Horvath2 .....	7
5.4 PhenoAge.....	8
5.5 GrimAge1 .....	8
5.6 GrimAge2 .....	8
5.7 DunedinPACE .....	9
6. Principal Component Versions of Clocks .....	10
6.1 Rationale .....	10
6.2 PC Clock Generation .....	10
6.3 Differences from Standard Clocks .....	10
7. Data Structure and Naming Conventions .....	12
7.1 File Formats .....	12
7.2 Variable Names, Labels, and Units .....	12
8. Interpretation Guidance .....	13
8.1 Recommended Covariates and Model Considerations .....	13
8.2 Limitations and Cautions .....	13
9. Example Code and Analyses .....	14
10. References .....	18
11. Contact and Support.....	20



# 1. Introduction

## 1.1 Overview of the Add Health Study

The National Longitudinal Study of Adolescent to Adult Health (Add Health) is a nationally representative sample of U.S. adolescents who were in grades 7-12 during the 1994-1995 school year. Using a complex, school-based cluster-sampling frame, researchers selected high school and feeder school pairs from 80 communities across the United States and drew a sex- and grade-stratified random sample of 20,745 adolescents for inclusion in the study. At Wave V, conducted between 2016 and 2018, participants were ages 33 to 43, with an average age of 38.

The study was designed to investigate how social, behavioral, and environmental factors interact with biological processes to influence health and developmental trajectories over the life course (Harris et al. 2019). Over the years, Add Health has collected a wealth of information from participants and their parents about demographic characteristics, familial structures, social relationships, health behaviors, cognition, physical and mental health status, medication usage, and health care access. Add Health also has collected anthropometric, cardiovascular, metabolic, renal, hepatic, inflammatory/immune, infectious, neurodegenerative, and multi-omic biomarkers from participants. In addition, Add Health has merged multilevel contextual data about the economic, school, neighborhood, policy, and environmental contexts in which the participants are embedded to the core survey and biological data at each wave. The Add Health dataset thereby provides researchers with rich opportunities to explore the causes and consequences of health status across multiple contextual domains as individuals age across the life course.

A central goal of Add Health Wave V was to expand the measurement of health-related biomarkers. To support this aim, an ancillary study (Harris and Aiello, MPIs R01MD013349) was developed to assess DNA methylation. With its extensive range of measures and longitudinal design, Add Health offers a unique opportunity to explore how social, environmental, and biological factors influence DNA methylation. As part of the DNA methylation data collection, several widely used biological clocks and related cellular measures were computed and made available to users, as detailed in this guide.

## 1.2 Rationale

Epigenetic clocks are biomarker algorithms that estimate aspects of biological aging, health risk, or physiological state based on DNA methylation levels at specific CpG sites. Add Health Wave V represents a transitional period into midlife- a key life stage when biological aging trajectories begin to diverge but before most chronic diseases (e.g., cardiovascular disease, diabetes, dementia) become prevalent. Measuring epigenetic age at Wave V provides critical information about early biological aging acceleration, offering insight into preclinical risk and disease development pathways. We computed several widely used DNA methylation-based clocks, including the original and updated Horvath multi-tissue estimators, PhenoAge, GrimAge (versions

1 and 2), and DunedinPACE, as well as principal component-based versions of each (except DunedinPACE) to enhance measurement reliability. In addition, we provide Houseman-estimated blood cell type proportions to facilitate adjustment for cellular heterogeneity in statistical analyses.

Together with the rich survey and biomarker data available in Add Health, these epigenetic clock measures enable researchers to investigate biological aging and its social, behavioral, and environmental determinants in a diverse, nationally representative cohort followed from adolescence into adulthood.

### 1.3 Purpose of This Document

This document describes the generation of epigenetic clocks and immune cell counts from DNA methylation data from participants of the National Longitudinal Study of Adolescent to Adult Health (Add Health). It provides detailed definitions, computation methods, data descriptions, and guidance for interpretation to ensure consistent and reproducible use of these measures. The content here complements existing Add Health documentation on Wave V biospecimen collection and methylation data.

## 2. Background

### 2.1 Add Health and the Methylation Ancillary Study

The methylation ancillary study was conducted using blood collected during the Add Health Wave V home exam and funded by R01MD013349 to MPIs Harris and Aiello. DNA methylation levels were measured using Illumina Infinium MethylationEPIC BeadChip (v1) technology, producing beta value matrices for downstream analysis. The methylation data are available through dbGaP (Add Health accession# phs001367.v3.p1). This document describes epigenetic clocks derived from these data that are disseminated through Add Health's restricted-use data use agreements.

### 2.2 Related Documentation

This document should be read in conjunction with the following Add Health technical documentation:

- [\*\*Blood Collection and Processing Protocols:\*\*](#) Sampling section describes biospecimen collection procedures, processing steps, storage conditions, and quality assurance methods Add Health dbGaP accession# phs001367.v2.p1 "Quality Control of Gene Expression Data, phd008756.1).
- [\*\*Methylation Beta Matrix Generation User Guide:\*\*](#) Details laboratory processing, methylation data preprocessing, probe filtering, normalization, and quality control steps used to generate the final beta matrices used as input for clock calculations (see Add Health dbGaP accession# phs001367.v3.p1).

### 3. Overview of Variables

Measure	Variable	Units	Description	PC Version	Reference
Identifier	aid	n/a	Unique identifier for Add Health participants	n/a	n/a
Chronological Age	age.at.w5	Years	Chronological age of participants at time of blood collection	n/a	n/a
Processing batch	batch	1/2a/2b/3/4	Batch numbers for processing epigenetic data	n/a	n/a
Houseman Immune Cell Estimates	cell_*	Proportion (0–1)	Estimates relative proportions of immune cell types (CD4T, CD8T, NK, B cells, monocytes, granulocytes)	No	Houseman et al., 2012
Horvath Clock v1	Horvath1	Years	Estimates chronological age from methylation levels across multiple tissues	Yes	Horvath, 2013
Horvath Clock v2	Horvath2	Years	Updated multi-tissue age estimator with improved accuracy across age ranges and population groups	Yes	Horvath et al., 2018
PhenoAge Clock	PhenoAge	Years	Estimates a “phenotypic age” based on methylation patterns linked to mortality-related clinical biomarkers	Yes	Levine et al., 2018
GrimAge v1	GrimAge1	Years	Predicts time-to-death and disease risk using methylation surrogates of plasma proteins and smoking pack-years	Yes	Lu et al., 2019
GrimAge v2	GrimAge2	Years	Updated GrimAge algorithm with refined CpG sets for improved mortality prediction	Yes	Lu et al., 2022
DunedinPACE	DunedinPACE	Years of biological aging/Calendar year	Estimates the pace of biological aging (rate of change in physiological integrity) using longitudinal biomarker data	No	Belsky et al., 2022



## 4. Data Processing Prior to Clock Calculation

### 4.1 Preprocessing of Beta Values

DNA methylation data for the epigenetic clocks were derived from Illumina Infinium Methylation EPIC BeadChip (v1) assays (Methylation Beta Matrix Generation User Guide). Raw intensity files were processed to generate beta values, representing the proportion of methylation at each CpG site on a 0–1 scale. Preprocessing steps included:

- Background correction using Illumina’s internal controls.
- Color channel adjustment to harmonize Type I and Type II probe designs.
- Normalization using functional normalization implemented in the **minfi** R package.
- Conversion of intensities to beta values, calculated as  $M / (M + U + 100)$ , where  $M$  and  $U$  are methylated and unmethylated signal intensities, respectively.

Only probes passing platform-specific detection p-value thresholds were retained prior to downstream analyses.

### 4.2 Quality Control Steps

Quality control (QC) procedures were applied to both probe-level and sample-level data to ensure high reliability of the methylation measures:

- **Sample-level QC**
  - Exclusion of samples with >1% of probes failing the detection p-value threshold ( $p > 0.01$ ). (n=0)
  - Removal of sex mismatches (n=1)
  - Removal of samples with identical genotypes (n=6)
  - Removal of technical artifacts (assay intensity failures, bisulfite conversion failure) (n=10)
- **Probe-level QC**
  - Removal of probes with detection p-value  $> 0.01$  in more than 5% of samples. (n=2479)
  - Removal of probes with known SNPs at the CpG. (n=30435)
  - Exclusion of cross-reactive probes identified in published lists (Pidsley et al., 2013). (n=43177)

### 4.3 Missing CpGs

Epigenetic clock algorithms require a defined set of CpG sites. When one or more required CpGs were missing after QC filtering, the following procedures were applied:

- CpG values missing for one or more samples were adjusted to NA. No imputation was done prior to using the MethylCipher, DunedinPACE or Horvath online calculator though these packages conduct imputation within their scope eg. methylCipher sets missing betas to 0, while DunedinPACE mean imputes values.







Refer to the following publications for greater detail (Thrush 2022, Belsky 2022, Horvath 2013). The proportion of probes present in the non-imputed data submitted to MethylCipher, DunedinPACE, and Horvath online calculator is shown below).

Clock	Present_Probes	Total_Probes	Percent_Present
Horvath1	334	353	94.6%
Horvath2	391	391	100.0%
PhenoAge	512	513	99.8%
GrimAge	1277	1331	95.9%
DunedinPACE	158	173	91.3%
PC Clocks	78445	78646	99.9%

## 5. Detailed Methods by Clock

### 5.1 Houseman Immune Cell Estimates

- **Definition:** Estimates relative proportions of major blood cell types including CD8 T cells, CD4 T cells, natural killer cells, B cells, monocytes, and granulocytes from methylation profiles.
- **Method:** Reference-based deconvolution using a set of CpGs informative of cell-type-specific methylation patterns.
- **Software/Parameters:** Implemented in the R package minfi (function estimateCellCounts), using the default reference dataset supplied by Houseman et al. (2012). No additional parameter tuning was applied.
- **Interpretation:** Values reflect distribution of cellular composition of the blood samples and may be used as covariates in epigenetic clock analyses to adjust for cellular heterogeneity.

Cells	Mean	SD	Min	Max	Median	Histogram
cell_CD4T	17.4%	6.2%	0.0%	50.5%	17.0%	
cell_CD8T	6.9%	4.6%	0.0%	37.4%	6.4%	
cell_Bcell	5.4%	2.9%	0.0%	25.5%	4.9%	
cell_NK	2.8%	3.3%	0.0%	40.3%	1.7%	
cell_Mono	5.6%	2.4%	0.0%	20.9%	5.4%	
cell_Gran	59.1%	10.2%	0.0%	97.7%	59.3%	

### 5.2 Horvath1

- **Definition:** The first pan-tissue DNA methylation age estimator, trained on methylation data from 51 tissue and cell types. This is the most commonly cited first generation clock.
- **Method:** Elastic net regression model using 353 CpGs to predict chronological age.
- **Software/Parameters:** Implemented using Steve Horvath Lab's publicly distributed DNAmAGE online calculator. Default coefficients were used. Sample type was annotated as whole blood. Sex and chronological age were also annotated.
- **Interpretation:** Output is in units of years, with values tracking chronological age but allowing acceleration/deceleration estimates when residualized on actual age.

### 5.3 Horvath2

- **Definition:** First-generation pan-tissue estimator designed for improved accuracy across diverse populations and extended age ranges.

- **Method:** Elastic net regression model based on ~3,500 CpGs selected from Illumina EPIC/450K arrays.
- **Software/Parameters:** Computed using the publicly available DNA methylation age calculator **methyLCipher** from Morgan Levine (2020). All default parameters were used. Sample type was annotated as whole blood. Sex and chronological age were also annotated. Default specifications applied
- **Interpretation:** Output in years. Provides improved calibration at older ages relative to Horvath1.

#### 5.4 PhenoAge

- **Definition:** Second-generation DNA methylation–based biomarker that predicts a composite “phenotypic age” derived from clinical blood chemistry markers associated with mortality.
- **Method:** Trained via elastic net regression to predict a weighted composite of 9 clinical biomarkers and chronological age. Final model uses 513 CpGs.
- **Interpretation:** Output in years. Higher PhenoAge values relative to chronological age indicate accelerated biological aging associated with morbidity and mortality risk.

#### 5.5 GrimAge1

- **Definition:** Second generation GrimAge, designed to predict time-to-death, coronary disease, and other healthspan outcomes by combining DNAm surrogates of plasma proteins and smoking pack-years, plus biological sex.
- **Method:** Elastic net regression model incorporating 1030 CpGs linked to surrogates of 7 plasma proteins and cumulative smoking exposure, plus biological sex.
- **Software/Parameters:** Implemented using Steve Horvath Lab’s publicly distributed DNAmAGE online calculator. Default coefficients were used. Sample type was annotated as whole blood. Sex and chronological age were also annotated.
- **Interpretation:** Output in years. Strong predictor of all-cause mortality, cardiovascular disease, and other age-related conditions.

#### 5.6 GrimAge2

- **Definition:** Updated GrimAge clock with refined CpG predictors and 2 additional plasma protein surrogates to enhance mortality prediction accuracy.
- **Method:** Similar structure to GrimAge1, using the same 1030 CpGs, but updated training set and revised CpG weights.
- **Software/Parameters:** Implemented using Horvath Lab’s publicly distributed DNAmAGE online calculator. Default coefficients were used. Sample type was annotated as whole blood. Sex and chronological age were also annotated.
- **Interpretation:** Output in years. Provides improved mortality and morbidity prediction over GrimAge1.

## 5.7 DunedinPACE

- **Definition:** A measure of the *pace of aging* that quantifies the rate of physiological decline (biological years of aging per chronological year), based on longitudinal biomarker data from the Dunedin Study cohort.
- **Method:** Elastic net regression of longitudinal organ-system integrity changes on DNA methylation, yielding an algorithm of 173 CpGs. An additional 19,827 CpGs are used for background normalization prior to clock calculation.
- **Software/Parameters:** Implemented using the **DunedinPACE R package** (Belsky et al., 2022). Default settings applied.
- **Interpretation:** Output is in standard deviation units with mean of 1.0 in the reference Dunedin cohort, representing “1 biological year per chronological year.” Values greater than 1.0 indicate accelerated aging while values less than 1.0 indicate decelerated aging. DunedinPACE is distinct from clock-based age predictors, focusing instead on the rate of aging at the time of assessment.

## 6. Principal Component Versions of Clocks

### 6.1 Rationale

Epigenetic clock estimates based on individual CpG sites can be sensitive to technical noise arising from probe specific variability, batch effects, or other assay-related artifacts. To address this, principal component (PC) based versions of the clocks have been developed. Instead of computing the clock value directly from the weighted sum of individual CpGs, these versions are calculated from principal components derived from the full methylation dataset. The PCs were then trained to predict the original clock value (for Horvath1, Horvath2, and GrimAge1) or the biomarker score used for the original training (for PhenoAge). This process yields, for each clock, a set of CpGs with weights to construct PCs, and a second set of weights applied to those PCs to construct the PC clock value.

This approach reduces the influence of measurement error at any single CpG site, thereby improving reliability and test–retest reproducibility. PC-based clocks are particularly advantageous in large epidemiologic studies where noise from technical variation may obscure biological signal.

### 6.2 PC Clock Generation

- **Input Data:** Beta values from the full set of high-quality CpGs retained after preprocessing and quality control.
- **Method:** Higgins-Chen and colleagues (2022) selected a set of 78,464 CpGs that overlapped between the 450K and EPIC arrays and were present in available DNAm datasets used to train the Horvath1, Horvath2, PhenoAge, and GrimAge clocks. For each clock, a separate dataset was used for training, but the same 78,464 CpGs from each of those datasets were input. First, PCs were generated from the 78,464 CpGs within a clock’s training dataset. These PCs were then subjected to feature selection via elastic net regression to predict either the standard epigenetic clock value for the same samples (for Horvath1, Horvath2, and GrimAge) or the biomarker-based phenotype score for the same samples (PhenoAge). This resulted in a reduced set of PCs with weights to calculate the ‘PC version’ of each clock, ranging from 121 retained PCs for Horvath1 to 1,936 retained PCs for GrimAge1.
- **Software/Parameters:** Computed in R using code that is publicly available on GitHub at: <https://github.com/MorganLevineLab/PC-Clocks>

### 6.3 Differences from Standard Clocks

- **Reliability:** PC-based clocks show higher test–retest reliability compared to standard versions, particularly in longitudinal or repeated-measures designs.
- **Interpretation:** Output units and biological meaning remain the same as the standard clock; the difference lies only in the underlying computational method.

- **Comparability:** Estimates from PC versions may differ slightly in scale or distribution compared to standard clocks. Researchers are encouraged to use PC versions for primary analyses when possible but may choose to use standard versions when aiming for comparability with prior studies that used non-PC versions of the clocks.
- **Availability:** Variable names for PC clocks follow the same prefix as their standard counterparts, with “PC\_” appended (e.g., PC\_Horvath1).

Clock	r2 ChronAge	Mean	Median	SD	Min	Max
Horvath1	0.18	39.20	39.2	4.34	15.50	62.30
Horvath2	0.34	38.30	38.3	3.46	20.40	60.50
PhenoAge	0.11	30.10	30.1	5.71	2.43	54.20
GrimAge1	0.16	42.60	41.9	4.58	27.30	62.30
GrimAge2	0.13	47.90	47.3	5.17	31.30	68.50
PCHorvath1	0.13	45.00	45	3.86	30.80	66.30
PCHorvath2	0.10	40.90	41	4.24	25.60	62.90
PCPhenoAge	0.11	41.10	40.8	5.45	20.90	66.50
PCGrimAge	0.19	54.20	53.7	3.80	43.10	71.90
age.at.w5	1.00	38.50	38.5	1.93	33.10	44.80

## 7. Data Structure and Naming Conventions

### 7.1 File Formats

Epigenetic clock variables are distributed in standard Add Health data release formats:

- SAS (.xpt)
- Delimited text (.csv)

All files contain identical variables and metadata, ensuring consistency across platforms.

Researchers should consult the Add Health Data User Guide for general guidance on importing and merging these files with other Add Health data products.

### 7.2 Variable Names, Labels, and Units

- Variable Naming Conventions
  - Each measure is prefixed with a short identifier.
  - Standard clocks: Horvath1, GrimAge2, PhenoAge, etc.
  - Principal component versions: PC appended (e.g., PCHorvath1, PCGrimAge2).
  - Houseman cell counts: cell\_CD8T, cell\_CD4T, cell\_NK, cell\_Bcell, cell\_Mono, cell\_Gran.
- Labels
  - Each variable is accompanied by a descriptive label indicating the measure, version, and units.
- Units
  - Most clocks are reported in years.
  - DunedinPACE is reported as years of biological aging per chronological year (dimensionless, mean ~1).
  - Houseman estimates are proportions between 0 and 1.



## 8. Interpretation Guidance

### 8.1 Recommended Covariates and Model Considerations

- **Chronological Age** - Include age in models when interpreting “age acceleration” (i.e., residuals of DNAm age regressed on chronological age).
- **Sex** - Biological sex may influence both methylation and health outcomes, making it an important covariate. Note that biological sex is also included in the calculation of GrimAge1 and GrimAge2.
- **Cell Composition** - Houseman-estimated immune cell proportions should be included in regression models to reduce confounding by leukocyte heterogeneity.
- **Technical Covariates** - Batch indicators, plate, or array position may be relevant, depending on analysis design.
- **Other Considerations** - Smoking and other lifestyle exposures strongly influence certain clocks (e.g., GrimAge). Adjusting for these variables depends on whether they are considered confounders or mechanisms of interest.

### 8.2 Limitations and Cautions

- **Population Generalizability** - Clock algorithms were trained in specific reference populations. While generally robust, performance in Add Health may differ by ancestry, age range, or health status of the training population.
- **Measurement Scale** - Clock values are continuous and probabilistic, not deterministic; individual-level predictions may deviate substantially from actual outcomes.
- **Interpretation of Differences** – There is no clinical consensus on how to interpret differences in DNAm age clocks. A one-year increase in DNAm age does not necessarily equal one calendar year of additional biological aging. Interpret differences as relative indicators of biological risk or aging pace rather than clinical differences in aging biology.
- **Comparability of Versions** - GrimAge1 vs. GrimAge2, and standard vs. PC clocks, may yield slightly different distributions. Researchers should select versions consistently within a study and document their choice.
- **Platform Effects** - Some clocks were originally developed on 450K arrays, while Add Health methylation data were generated on EPIC arrays. Small discrepancies may arise from probe coverage differences.

## 9. Example Code and Analyses

```
# libraries
pacman::p_load(data.table,tidyverse,methylCIPHER,DunedinPACE,Hmisc)

# Invalid cpgs from detectionP
inv = fread('~/.epigenetics/processRawData/betas/postQCbetas/invalid_cpgs.csv')

# Cross reactive probes from Pidsley
pids = fread('~/.epigenetics/processRawData/betas/postQCbetas/pidsley-2016-S1.csv')

# Read in betas
b1 = fread('~/.epigenetics/processRawData/betas/beta.batch1.csv')
b2a = fread('~/.epigenetics/processRawData/betas/beta.batch2a.csv')
b2b = fread('~/.epigenetics/processRawData/betas/beta.batch2b.csv')
b3 = fread('~/.epigenetics/processRawData/betas/beta.batch3.csv')
b4 = fread('~/.epigenetics/processRawData/betas/beta.batch4.csv')

# Remove probes from betas and round
b1 = b1 %>% filter(!cg %in% c(pids$V1,inv$inv)) %>%
mutate(across(where(is.numeric),round,digits=6))
b2a = b2a %>% filter(!cg %in% c(pids$V1,inv$inv)) %>%
mutate(across(where(is.numeric),round,digits=6))
b2b = b2b %>% filter(!cg %in% c(pids$V1,inv$inv)) %>%
mutate(across(where(is.numeric),round,digits=6))
b3 = b3 %>% filter(!cg %in% c(pids$V1,inv$inv)) %>%
mutate(across(where(is.numeric),round,digits=6))
b4 = b4 %>% filter(!V1 %in% c(pids$V1,inv$inv)) %>%
mutate(across(where(is.numeric),round,digits=6)) %>% dplyr::rename(cg=V1)
colnames(b4)[2:ncol(b4)] = substr(colnames(b4)[2:ncol(b4)],2,21)

# Combine betas
b = b1 %>% full_join(b2a,by='cg') %>% full_join(b2b,by='cg') %>% full_join(b3,by='cg')
%>% full_join(b4,by='cg')

#####DNAMAge Input File#####
# Make dnamage input
dnamage = fread('~/.epigenetics/makeClocks/dnamage/datMiniAnnotation4_fixed.csv') %>%
dplyr::select(Name) %>%
dplyr::rename(cg=Name)

#####Annotation File#####
### Collect data from all 5 waves

#### define path
path = '/ifs/sec/cpc/addhealth/addhealthcontractdata/core/'
setwd(path)

#### load survey data and home.exam demographics
w1 = paste0(path,'wave1/allwave1.xpt')
w2 = paste0(path,'wave2/wave2.xpt')
w3 = paste0(path,'wave3/wave3.xpt')
w4 = paste0(path,'wave4/wave4.xpt')
w5 = paste0(path,'wave5/WAVE5.xpt')
home.exam =
sasxport.get('/ifs/sec/cpc/addhealth/addhealthcontractdata/wave_5_biomarkers/bdemo5.xp
t')
```

```
##### import files
w1 = sasxport.get(w1)
w2 = sasxport.get(w2)
w3 = sasxport.get(w3)
w4 = sasxport.get(w4)
w5 = sasxport.get(w5)

##### fix aid data type
w1$aid = as.character(w1$aid)
w2$aid = as.character(w2$aid)
w3$aid = as.character(w3$aid)
w4$aid = as.character(w4$aid)
w5$aid = as.character(w5$aid)
home.exam$aid = as.character(home.exam$aid)

##### combine into one dataframe
survey = w1 %>%
  left_join(w2,by='aid')%>%
  left_join(w3,by='aid')%>%
  left_join(w4,by='aid')%>%
  left_join(w5,by='aid') %>%
  left_join(home.exam,by='aid')

### age at w5 home exam
age.at.w5.home.exam = survey %>%

# make age at wave5 home exam
mutate(age.at.w5.home.exam = as.numeric(
  as.Date(paste0('01/',h5exmon,'/',h5exyear),format='%d/%m/%Y') -
  as.Date(paste0('01/',h5od1m,'/',h5od1y),format='%d/%m/%Y'))/365)%>%

mutate(age.at.w5.home.exam.backup = as.numeric(
  as.Date(paste0('01/',h5exmon,'/',h5exyear),format='%d/%m/%Y') -
  as.Date(paste0('01/',h1gi1m,'/',h1gi1y),format='%d/%m/%y'))/365)%>%

mutate(age.at.w5.home.exam =
ifelse(is.na(age.at.w5.home.exam),age.at.w5.home.exam.backup,age.at.w5.home.exam)) %>%

# fix aid data type
mutate(aid = as.character(aid)) %>%

mutate(dateofbirth =
  as.Date(paste0('01/',h1gi1m,'/',h1gi1y),format='%d/%m/%y'))%>%

# select variables
dplyr::select(aid,age.at.w5.home.exam,dateofbirth)

#### sex
sex = survey %>%

#select variables
dplyr::select(aid,h5od2a,bio.sex,bio.sex4,bio.sex2,bio.sex3) %>%

#fix data type
```

```

mutate(aid = as.character(aid)) %>%

#rename
mutate(h5od2a = ifelse(h5od2a==6,bio.sex,h5od2a)) %>%
  mutate(h5od2a = ifelse(is.na(h5od2a),bio.sex,h5od2a)) %>%
  mutate(h5od2a = ifelse(h5od2a==6,bio.sex4,h5od2a)) %>%
  mutate(h5od2a = ifelse(is.na(h5od2a),bio.sex4,h5od2a)) %>%

mutate(sex = as.factor(ifelse(h5od2a==1,'male',
                             ifelse(h5od2a==2,'female',NA))))%>%

#select variables
dplyr::select(aid,sex)

# make crosswalk of aid to rownames
cw = fread('~/.epigenetics/processRawData/manifest/manifest.allBatches.csv') %>%
  mutate(coordinates = paste0(ChipBarcode,'_',ChipPos))%>%
  dplyr::select(coordinates,newAid,FinalQCStatus)

# make annotation file for online calculator
annotation = data.frame(colnames(b)[2:4799])
names(annotation) = 'coordinates'
annotation = annotation %>%

# merge annotation with crosswalk
left_join(cw,by='coordinates') %>%

# collect seed aid and merge with age.at.home.exam
mutate(seedAid = ifelse(endsWith(newAid,'R'),substr(newAid,1,8),newAid)) %>%
left_join(age.at.w5.home.exam,by=c('seedAid'='aid'))%>%

# merge with sex
left_join(sex,by=c('seedAid'='aid')) %>%
mutate(Sex=ifelse(sex=='male',0,1))%>%

# rename variables
dplyr::rename(Sample_Name=coordinates,Age=age.at.w5.home.exam)%>%

# make tissue type
mutate(Tissue='Whole Blood') %>%

# order annotation file
dplyr::select(Sample_Name,Age,Sex,Tissue,newAid)

fwrite(annotation,'~/epigenetics/makeClocks/dnamage/post_detectionp/postqc.annotation.
aid.csv',quote=F,row.names=F,sep=',')
annotation=annotation%>%dplyr::select(-newAid)
fwrite(annotation,'~/epigenetics/makeClocks/dnamage/post_detectionp/postqc.annotation.
csv',quote=F,row.names=F,sep=',')

# only keep cpgs in datMiniAnnotation4_fixed.csv file
b.dnamage = b %>%
filter(cg %in% dnamage$cg)

# write dnamage input

```

```

fwrite(b.dname, '~/epigenetics/makeClocks/dname/post_detectionp/postqc.dnameinput
.betas.csv', quote=F, row.names=F, sep=', ')

#####MethylCIPHER#####

# run methylcipher
# load clocks
clocks =
c("calcHorvath1",
"calcHorvath2",
"calcPhenoAge")

# load cpGs
Horvath1_CpGs = methylCIPHER::Horvath1_CpGs
Horvath2_CpGs = methylCIPHER::Horvath2_CpGs
PhenoAge_CpGs = methylCIPHER::PhenoAge_CpGs

# invert betas
batches = b %>% column_to_rownames('cg') %>% as.matrix %>% t

# generate outcomes
outcome = data.frame(matrix(ncol=1, nrow=nrow(batches)))

# add coordinates to outcomes
outcome$coordinates = row.names(batches)

# remove empty columns
outcome[,1] = NULL

# generate outcomes
outcome = methylCIPHER::calcUserClocks(clocks, batches, outcome, imputation=F) %>%

# add newAid
left_join(cw, by='coordinates')

# save to file
fwrite(outcome, '~/epigenetics/makeClocks/postqc.methylcipher.outcome.csv', quote=F, row.
names=F, sep=', ')

#####DunedinPACE#####
# make dunedinPACE
betas = b %>% column_to_rownames('cg')
dunedin = data.frame(PACEProjector(betas)) %>%
rownames_to_column('coordinates') %>%
left_join(cw, by='coordinates') %>%
mutate(DunedinPACEZ = as.numeric(scale(DunedinPACE)))

# write DunedinPACE
fwrite(dunedin, '~/epigenetics/makeClocks/dunedin/postqc.dunedin.csv', quote=F, row.names
=F, sep=', ')

```

## 10. References

- Belsky, D. W., Caspi, A., Arseneault, L., Baccarelli, A., Corcoran, D. L., Gao, X., ... Moffitt, T. E. (2020). Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. *ELife*, 9, e54870.
- Belsky, D. W., Caspi, A., Corcoran, D. L., Sugden, K., Poulton, R., Arseneault, L., ... Moffitt, T. E. (2022). DunedinPACE, a DNA methylation biomarker of the pace of aging. *ELife*, 11, e73420.
- Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sada, S. V., ... Zhang, K. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular Cell*, 49(2), 359–367.
- Harris, K. M. (2010). Design features of Add Health. Retrieved from <http://www.cpc.unc.edu/projects/addhealth/data/guides/design%20paper%20WI-IV.pdf>.
- Harris, K. M., Levitt, B., Gaydos, L., Martin, C., Meyer, J. M., Mishra, A. A., Kelly, A. L., & Aiello, A. E. (2024). Sociodemographic and lifestyle factors and epigenetic aging in US young adults: NIMHD Social Epigenomics Program. *JAMA Network Open*, 7(7), e2427889.
- Harris K.M., Halpern C.T., Whitel E., Hussey J.M., Killea-Jones L., Tabor J., Dean, S.C.. Cohort Profile: The National Longitudinal Study of Adolescent to Adult Health (Add Health). *International Journal of Epidemiology* 2019;48(5):1415-1425.
- Higgins-Chen, A. T., Thrush, K. L., Wang, Y., Minter, C. J., Kuo, P. L., Wang, M., Niimi, P., Sturm, G., Lin, J., Moore, A. Z., Bandinelli, S., Vinkers, C. H., Vermetten, E., Rutten, B. P. F., Geuze, E., Okhuijsen-Pfeifer, C., van der Horst, M. Z., Schreier, S., Gutwinski, S., Luykx, J. J., Picard, M., Ferrucci, L., Crimmins, E. M., Boks, M. P., Hägg, S., Hu-Seliger, T. T., & Levine, M. E. (2022). A computational solution for bolstering reliability of epigenetic clocks: Implications for clinical trials and longitudinal tracking. *Nature Aging*, 2(7), 644–661.
- Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10), 3156.
- Horvath, S., Oshima, J., Martin, G. M., Lu, A. T., Quach, A., Cohen, H., ... Raj, K. (2018). Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging*, 10(7), 1758–1775.
- Houseman, E. A., Kile, M. L., Christiani, D. C., Ince, T. A., Kelsey, K. T., & Marsit, C. J. (2016). Reference-free deconvolution of DNA methylation data and mediation by cell composition effects. *BMC Bioinformatics*, 17, 259.
- Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., ... Horvath, S. (2018). An epigenetic biomarker of aging for lifespan and healthspan. *Aging*, 10(4), 573–591.

Lu, A. T., Binder, A. M., Zhang, J., Yan, Q., Reiner, A. P., Cox, S. R., ... & Horvath, S. (2022). DNA methylation GrimAge version 2. *Aging*, 14(23), 9484.

Lu, A.T., Quach, A., Wilson, J. G., Reiner, A. P., Aviv, A., Raj, K., ... Horvath, S. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging*, 11(2), 303–327.

Pidsley, R., Zotenko, E., Peters, T. J., Lawrence, M. G., Risbridger, G. P., Molloy, P., Van Djik, S., Muhlhausler, B., Stirzaker, C., & Clark, S. J. (2016). Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biology*, 17(1), 208.

Thrush, Kyra, Higgins-Chen, Albert Liu, Zuyun, Levine, Morgan. (2022). R methylCIPHER: A Methylation Clock Investigational Package for Hypothesis-Driven Evaluation & Research. 10.1101/2022.07.13.499978.

Whitsel EA, Angel R, O'Hara R, Qu L, Carrier K, Harris K. Add Health Wave V Documentation: Updated Measures of Inflammation and Immune Function, 2024; <https://doi.org/10.17615/7c3j-gd57>.

## 11. Contact and Support

Add Health is a program project directed by Kathleen Mullan Harris and designed by J. Richard Udry, Peter S. Bearman, and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill, and funded by grant P01-HD31921 from the Eunice Kennedy Shriver National Institute of Child Health and Human Development, with cooperative funding from 23 other federal agencies and foundations.

Information on how to obtain the Add Health data files is available on the Add Health website (<http://www.cpc.unc.edu/addhealth>).