Understanding Society: Waves 2-3 Nurse Health Assessment

‘Epigenetic ageing algorithms’ derived from DNA methylation, 2010-2012

USER GUIDE

Version 3, June 2025

Yanchun Bao[1,2]
Tyler Gorrie-Stone[3]
Meena Kumari[1]

[1] Institute for Social and Economic Research, University of Essex
[2] Department of Mathematical Sciences, University of Essex
[3] School of Life Sciences, University of Essex

Understanding Society

UKRI Economic and Social Research Council

Contents

1. Introduction

Understanding Society is a longitudinal household panel survey that started in 2009 and attempts to interview the core sample members at approximately one year intervals, as long as they are living in the UK. Each sets of interviews are referred to as waves. After the 2nd and 3rd waves sub-samples of the Study consented to nurse visits, blood and DNA extraction. To learn more about the survey please visit the About the Study webpage and the Main Survey User Guide. To learn more about the nurse visits and the health and biomarker data collected please visit the Health and biomarkers webpage.

In this document we give a brief introduction to seven epigenetic 'ageing algorithms' (clocks) constructed using *Understanding Society*, UK household longitudinal study (UKHLS) DNA methylation (DNAm) data. Epigenetic ageing algorithms, constructed based on a set of CpG sites whose DNA methylation levels are associated to chronological age or ageing-related health outcomes, have attracted a lot of research interests for their potential to quantify rates of biological ageing [1]. The difference between a person's chronological age and epigenetic age calculated by these 'ageing algorithms' has been used as an indicator of whether an individual is ageing faster or slower biologically than expected given their actual age. This difference may be related to his/her life circumstances and environments and has been associated with health and mortality.

These 'ageing algorithms' are referred to by the first author of the publication in which they are described or the name given to them in the publication. Five epigenetic ageing algorithms, Horvath 2013, Hannum, PhenoAge, Horvath skin and blood and Lin ageing algorithm, have been constructed using *Understanding Society* DNAm data with function "*agep()*" in R package "*wateRmelon*" [2]. We also provide two algorithms that were created as measures of the pace or speed of ageing. DunedinPoAm, has been constructed with function "*PoAmProjector()*" in R package "DuneDinPoAm38" (https://github.com/DLCorcoran/DunedinPoAm38) and DunedinPACE, has been constructed with function "*PACEProjector()*" in R package "DuneDinPACE" (https://github.com/DLCorcoran/DunedinPACE).

1.1 Data collection

Between 2010 and 2012, eligible adult participants of Wave 2, (general population sample (GPS)) and Wave 3 (British Household Panel Survey (BHPS)) of *Understanding Society* received a health assessment visit from a registered nurse. Blood samples were also taken at these visits and DNA was extracted from these samples. DNA methylation profiles were obtained from DNA extracted from whole blood from 3,654 eligible individuals who had consented to both blood sampling and genetic analysis during 2010-2012. There were 1,425 samples from individuals that took part in Wave 2 and another 2,229 from Wave 3. Details of eligibility criteria can be found in the *Understanding Society* Biomarker user guide and

glossary [3]. Eligibility requirements for genetic analyses meant that the epigenetic samples were restricted to participants of white ethnicity.

1.2 Understanding Society DNA methylation samples

Methylation profiling was conducted in two batches, consisting of n=1,174 (58.4% female, mean age 58.0, range: 28~98) measurements in 2017 from a subset of Wave 3 BHPS samples and another n=2,480 (54.2% female, mean age 50.5, range: 16~83) measurements made in 2020, from the rest of Wave 3 BHPS samples and a random subset from Wave 2 GPS samples. All measurements are from White/European participants with education distribution as degree (21.5%), other higher degree (13.5%), A-level (18.8%), GCSE (22.8%), other qualification (10.4%) and no qualification (13.0%). The sample (BHPS or GPS) can be identified using the variable "**hhorig**" and the batch can be identified using the variable "**Sample_measurement_file**".

1.3 QC-steps of methylation

Using the Illumina Methylation EPIC BeadChip, over 850,000 Methylation sites across the genome have been measured. Quality control steps, including outlier removal, filtering poor-quality probes and quantile normalisation have been pre-processed. R package "wateRmelon" [2] and "bigmelon" [4] were used to perform the quality control and normalisation of data.

The 2017 and 2020 datasets were analysed separately using the steps:

- Data-set read into R (2017: n = 1,193, # probe before QC = 866,895; 2020: n = 2,536, # probe before QC = 866,150)
- Outliers identified and removed using `wateRmelon::outlyx`
- Low quality samples (<85% bisulfite conversion) identified and removed using `wateRmelon::bscon`.
- Data was normalised using `wateRmelon::dasen`, difference between normalised and raw data per sample estimated using `wateRmelon::qual`. Samples were removed on the basis of having an Root mean square difference and standard deviation of difference > 0.05
- After removal of outlying/poor quality samples from raw data, data was subjected to `wateRmelon::pfilter` and then renormalised using `wateRmelon::dasen`.

These steps result in 1,174 Samples and 857,071 probes remaining for the 2017 dataset and 2,480 Samples and 860,950 remaining in the 2020 dataset.

2. Individual ageing algorithms

2.1 Horvath 2013

Horvath [5] developed a multi-tissue epigenetic ageing algorithm using 8,000 samples from 82 Illumina DNA methylation array datasets. 353 CpG sites had been used in [5] to construct an ageing algorithm which demonstrated the significant age acceleration linked with 20 cancer types. We used 333 out of 353 CpG sites to generate the *Understanding Society* Horvath 2013 ageing algorithm and the information of missing probes can be found in supplementary, part 1.

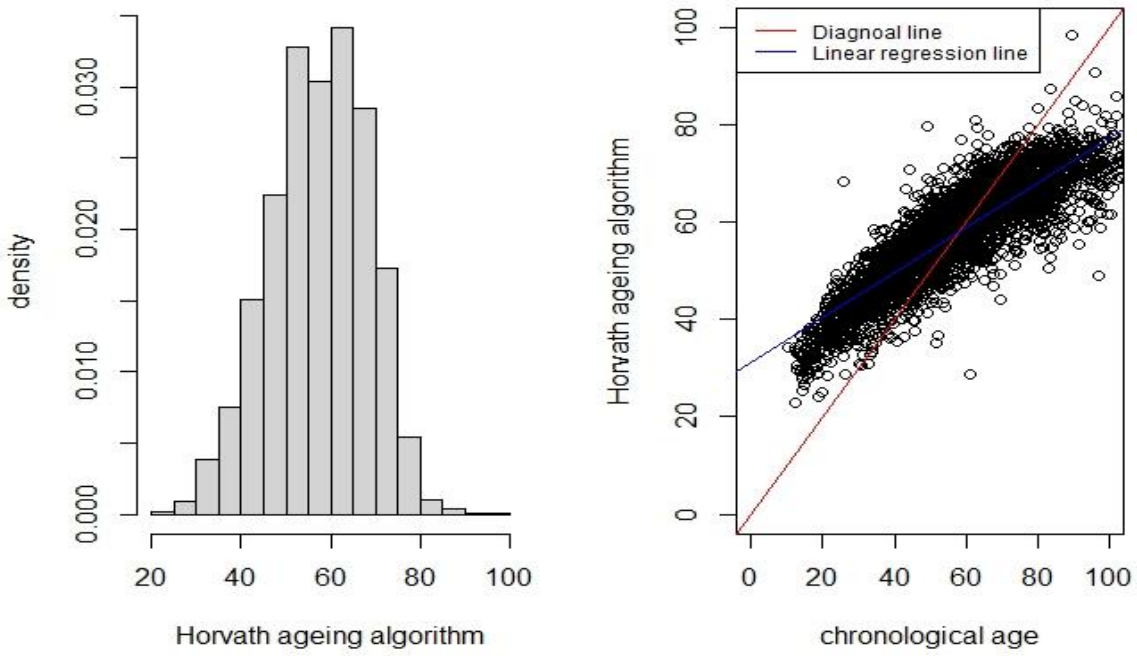
Table 1: Distribution of *Understanding Society* Horvath 2013 ageing algorithm

Horvath 2013	Sample size	Min	Max	Range	Median	Mean	Std	r*
2017 samples	1,174	28.8	98.6	69.8	57.9	57.7	10.4	0.90
2020 samples	2,480	23.1	90.8	67.7	57.7	57.1	11.0	0.94
Combined	3,654	23.1	98.6	75.5	57.7	57.3	10.8	0.91

* The correlation coefficient of the ageing algorithm with chronological age

Figure 1: Histogram of Horvath 2013 ageing algorithm (left) and

Scatter plot of chronological age with Horvath 2013 ageing algorithm (right)



2.2 Hannum

Hannum et.al [6] used a quantitative ageing model based on CpG sites measured from Illumina 450k array (HumanMethylation450K BeadChip) of whole blood samples of 656 individuals. 71 CpG sites were identified from their model and used to construct the Hannum ageing algorithm, where only 64 CpG sites are available with EPIC Array of *Understanding Society* samples and the information about missing CpG sites can be found in supplementary part 2.

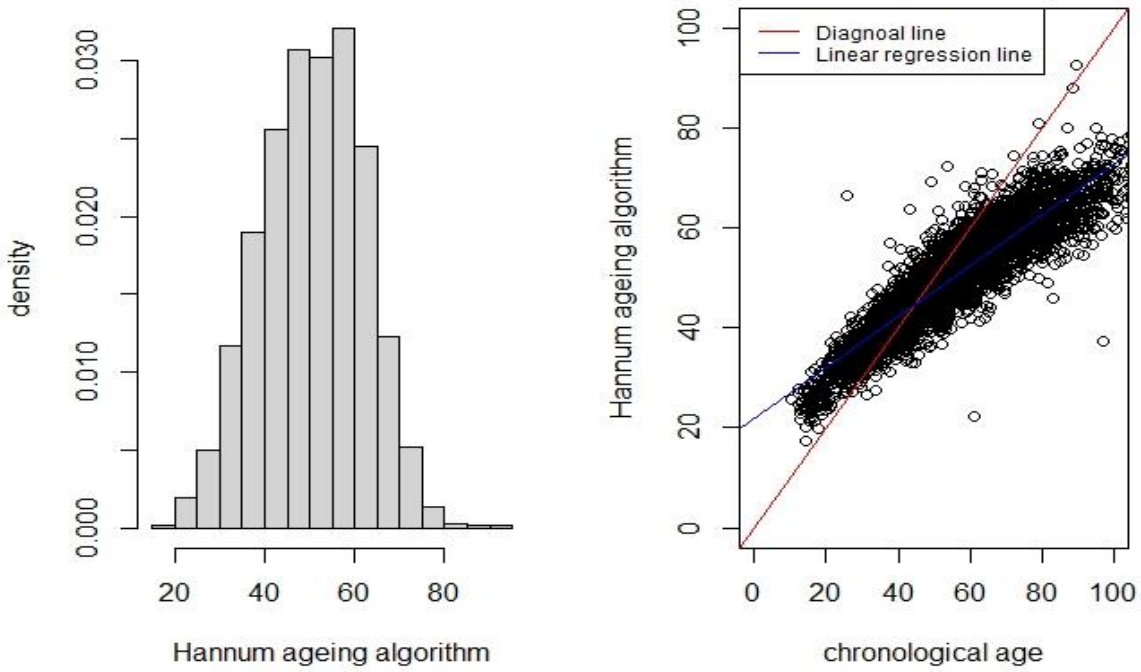
Table 2: Distribution of *Understanding Society* Hannum ageing algorithm

Hannum	Sample size	Min	Max	Range	Median	Mean	Std	r*
2017 samples	1,174	22.3	92.6	70.3	52.7	53.0	11.0	0.93
2020 samples	2,480	17.3	81.2	63.9	50.1	49.5	11.3	0.95
Combined	3,654	17.3	92.6	75.3	50.9	50.6	11.3	0.94

* The correlation coefficient of the ageing algorithm with chronological age

Figure 2: Histogram of Hannum ageing algorithm (left) and

Scatter plot of chronological age with Hannum ageing algorithm (right)



2.3 PhenoAge

Levine et.al [7] developed an ageing algorithm under the hypothesis that clinical measures based phenotypic age can identify novel CpGs and facilitate the power of epigenetic biomarker of ageing. 513 CpGs had been identified in [7] where 512 CpGs were used with the *Understanding Society* sample to generate the PhenoAge ageing algorithm (information about the missing CpGs sites can be found in supplementary. Part 3).

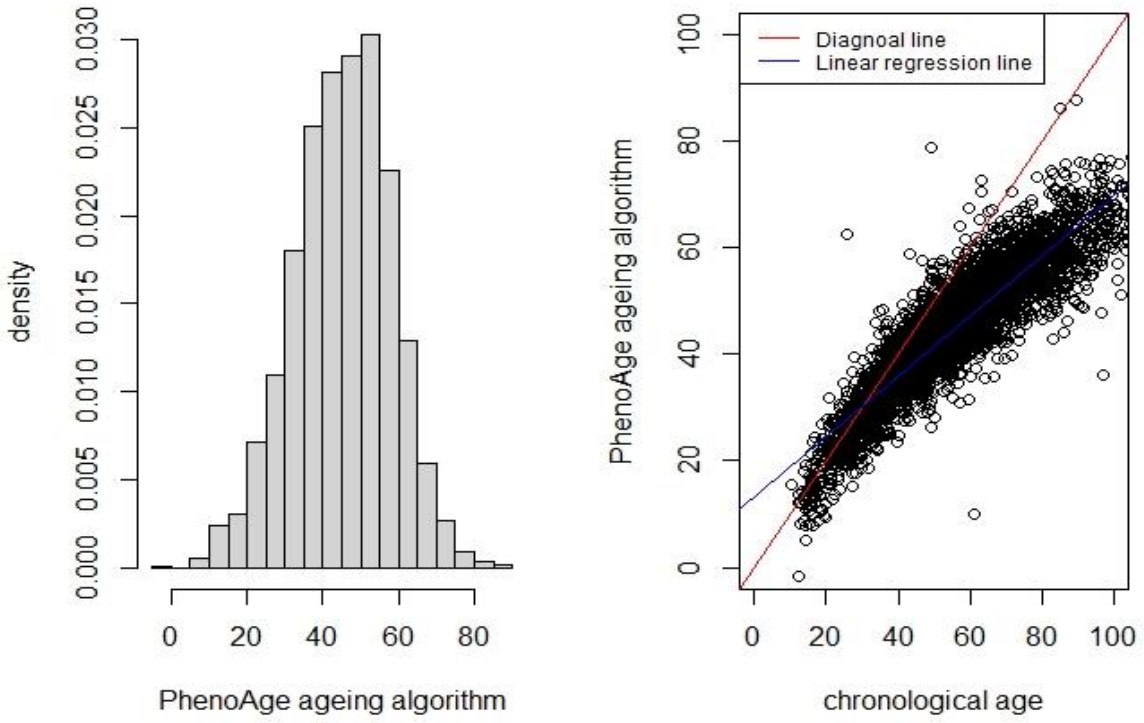
Table 3: Distribution of *Understanding Society* PhenoAge ageing algorithm

PhenoAge	Sample size	Min	Max	Range	Median	Mean	Std	r*
2017 samples	1,174	9.9	87.9	78.0	48.8	49.0	12.1	0.87
2020 samples	2,480	-1.6	76.6	78.2	44.4	43.4	12.6	0.92
Combined	3,654	-1.6	87.9	89.5	45.7	45.2	12.7	0.90

* The correlation coefficient of the ageing algorithm with chronological age

Figure 3: Histogram of PhenoAge ageing algorithm (left) and

Scatter plot of chronological age with PhenoAge ageing algorithm (right)



2.4 Horvath skin & blood

Horvath et.al [8] developed a DNA methylation ageing algorithm based on 391 CpGs from human fibroblasts, keratinocytes, buccal cells, lymphoblastoid cells, skin, blood and saliva samples. This age estimator is referred to as the skin & blood ageing algorithm. With EPIC array, all 391 CpGs were used to construct Horvath skin & blood ageing algorithm with *Understanding Society* samples.

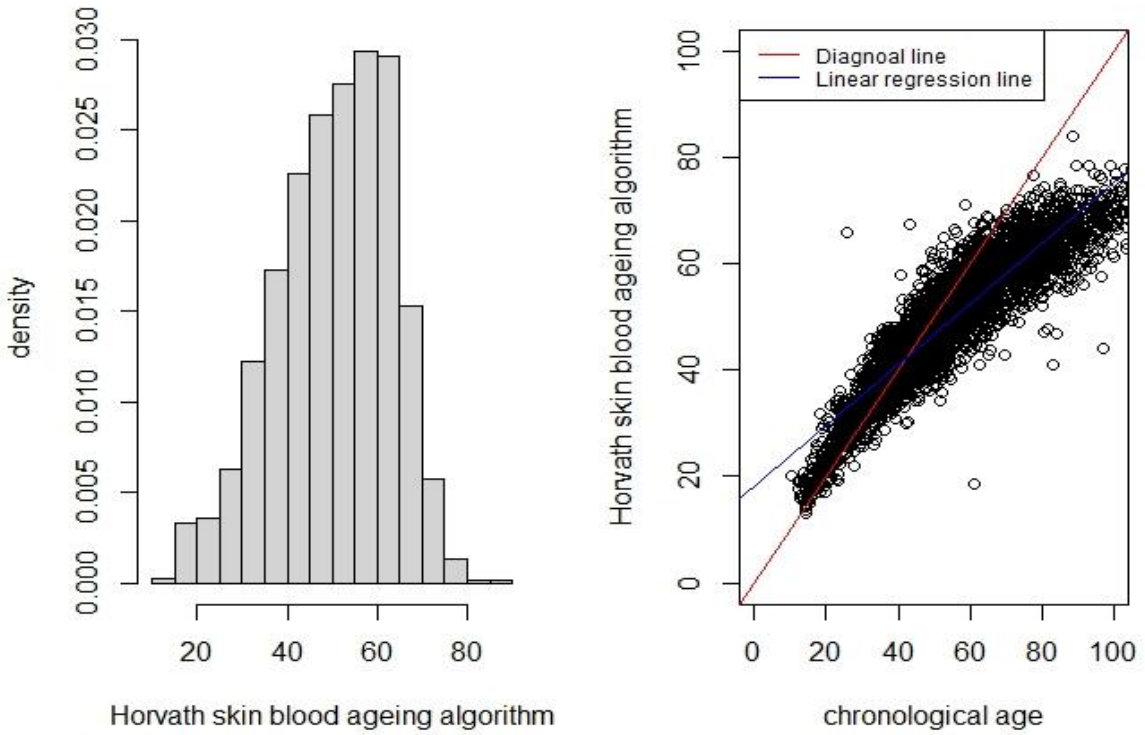
Table 4: Distribution of *Understanding Society* Horvath skin & blood ageing algorithm

Horvath skin & blood	Sample size	Min	Max	Range	Median	Mean	Std	r*
2017 samples	1,174	18.7	89.5	70.8	54.7	53.8	11.8	0.96
2020 samples	2,480	13.2	78.4	65.2	50.3	48.8	12.9	0.97
Combined	3,654	13.2	89.5	76.3	51.6	50.4	12.8	0.97

* The correlation coefficient of the ageing algorithm with chronological age

Figure 4: Histogram of Horvath skin & blood ageing algorithm (left) and

Scatter plot of chronological age with Horvath skin & blood ageing algorithm (right)



2.5 Lin

Lin and colleagues [9-11] developed an ageing predictor with 5,621 DNA methylation profiles of 25 cancer types. Among 99 CpG sites that used in [9], 2 are missing in EPIC Array and 97 are used to generate this ageing algorithm for *Understanding Society* samples (missing CpG sites information are given in supplementary, part 4).

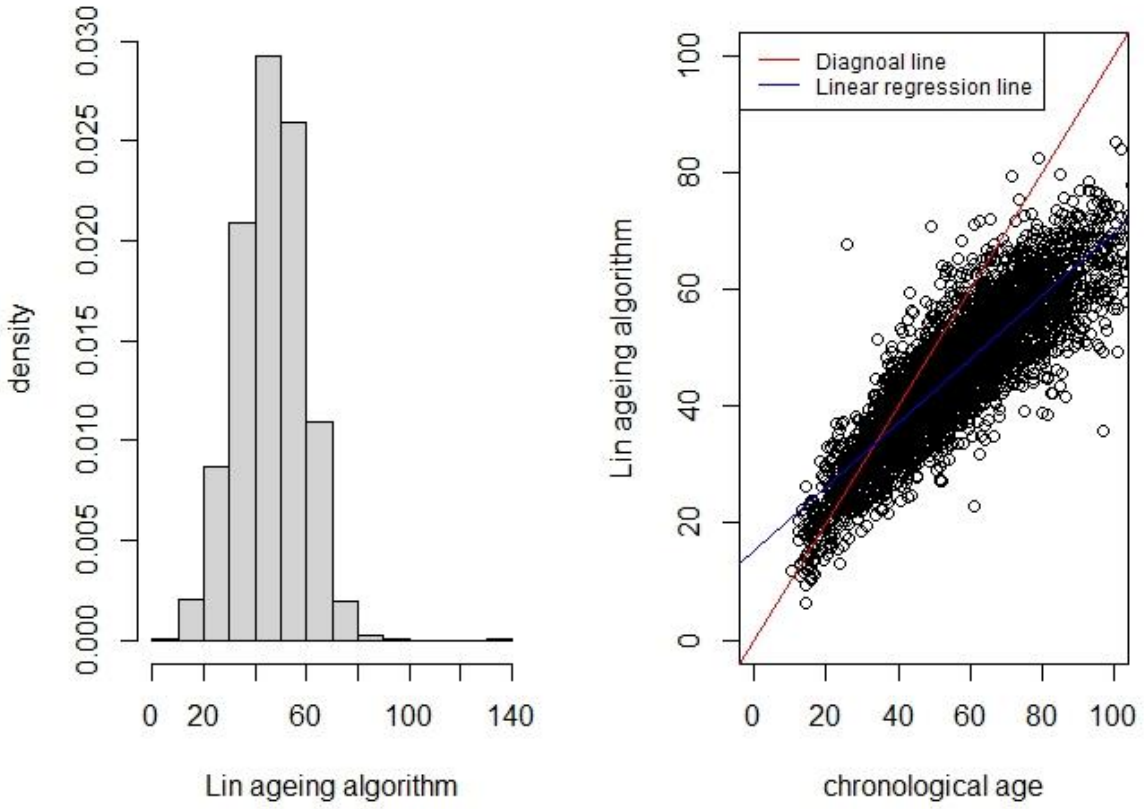
Table 5: Distribution of *Understanding Society* Lin ageing algorithm

Lin	Sample size	Min	Max	Range	Median	Mean	Std	r*
2017 samples	1,174	21.6	130.1	108.5	50.3	50.9	12.4	0.89
2020 samples	2,480	6.3	96.9	90.3	44.5	43.9	12.3	0.92
Combined	3,654	6.3	130.1	125.8	46.4	46.1	12.8	0.92

* The correlation coefficient of the ageing algorithm with chronological age

Figure 5: Histogram of Lin ageing algorithm (left) and

Scatter plot of chronological age with Lin ageing algorithm (right)



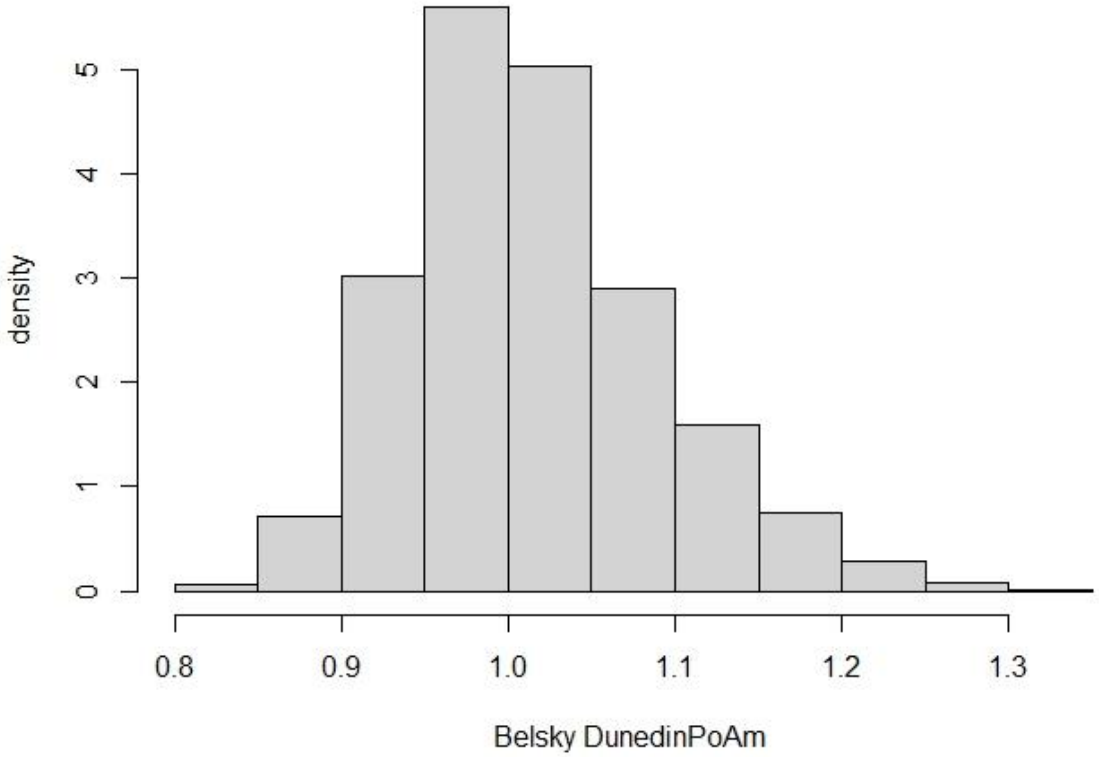
2.6 Belsky: DunedinPoAm38

Belsky and colleagues [12] developed an ageing algorithm, called DunedinPoAm38, to measure the biological 'pace of ageing'. Among 46 CpG sites that were used in [12], 1 is missing in the EPIC Array from *Understanding Society* samples (missing CpG sites information are given in supplementary, part 5). DunedinPoAm38 in [12] is estimated as a measurement of biological ageing years per chronological year (years/ chronological year) and the distribution is given in Table 6. In Figure 6, we plot the histogram of DunedinPoAm38.

Table 6: Distribution of *Understanding Society* Belsky DunedinPoAm

DunedinPoAm38	Sample size	Min	Max	Range	Median	Mean	Std
2017 samples	1,174	0.83	1.29	0.44	1.00	1.01	0.07
2020 samples	2,480	0.83	1.33	0.50	1.02	1.03	0.07
Combined	3,654	0.83	1.33	0.50	1.01	1.01	0.07

Figure 6: Histogram of Belsky DunedinPoAm



2.7 Belsky: DunedinPACE

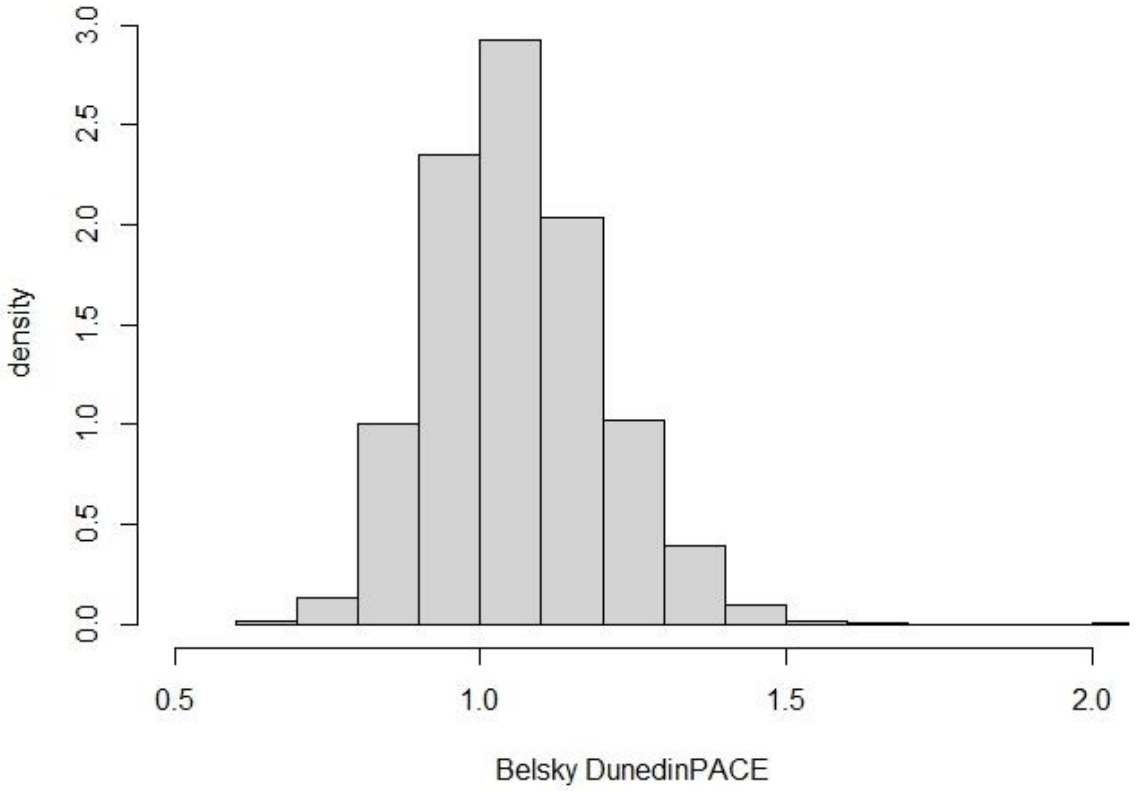
Belsky and colleagues [13] developed a further ageing algorithm, called DunedinPACE, which is a next-generation DNA-methylation biomarker of 'Pace of Ageing' that represents speed of ageing. Among 173 CpG sites that were used in [13], 1 is missing in the EPIC Array from *Understanding Society* samples (missing CpG sites information are given in supplementary, part 6). Similar to DunedinPoAm in 2.6, DunedinPACE in [13] is estimated as a measurement of biological ageing years per chronological year (years/chronological year) and the distribution is given in Table 7. In Figure 7, we show the histogram of the DunedinPACE.

Table 7: Distribution of *Understanding Society* Belsky DunedinPACE

DunedinPACE	Sample size	Min	Max	Range	Median	Mean	Std
2017 samples	1,174	0.66	2.00	1.34	1.07	1.07	0.14
2020 samples	2,480	0.63	1.68	1.05	1.04	1.05	0.14
Combined	3,654	0.63	2.00	1.37	1.05	1.06	0.14

* The correlation coefficient of the ageing algorithm with chronological age

Figure 7: Histogram of Belsky DunedinPACE



3. Data and Analysis Guidance

3.1 List of variables

The data is stored in the datafile, **xepigen_clocks_ns**. The list of variables and a brief description (variable label) in this datafile are in the table below.

Variable	Label
pidp	cross-wave person identifier (public release)
pid	personal identifier (BHPS cohort)
b_hidp	household identifier (public release)
b_pno	person number
b_splitnum	split number, 1=first, 2=second
c_hidp	household identifier (public release)
c_pno	person number
c_splitnum	split number, 1=first, 2=second
hhorig	sample origin BHPS or Understanding Society
wave	health assessment wave
DNAme_horvath_dv	Horvath - Epigenetic (DNA methylation) ageing algorithm
DNAme_hannum_dv	Hannum - Epigenetic (DNA methylation) ageing algorithm
DNAme_phenoage_dv	PhenoAge - Epigenetic (DNA methylation) ageing algorithm
DNAme_horvath_sb_dv	Horvath skin and blood - Epigenetic (DNA methylation) ageing algorithm
DNAme_lin_dv	Lin - Epigenetic (DNA methylation) ageing algorithm
DNAme_belsky_dunedinpoam_dv	Belsky_DunedinPoAm - Epigenetic (DNA methylation) speed of ageing
DNAme_belsky_dunedinpace_dv	Belsky_DunedinPACE - Epigenetic (DNA methylation) speed of ageing
Sample_measurement_file	Measurement sample year (2017, 2020)
barcode	Barcode
cd8t	CD8T
cd4t	CD4T
nk	NK
bcell	Bcell
mono	Mono
gran	Gran

3.2 Technical covariates

We suggest that the following technical covariates should be included in the analysis of the *Understanding Society* epigenetic ageing algorithm, including **barcode** (to account for the systematic effect of different batches in the experiment) and various cell composition estimates (**cd8t**, **cd4t**, **nk**, **bcell**, **mono**, **gran**). The cell composition estimates were calculated based on

Houseman reference-based algorithm implemented in the "*estimateCellCounts()*" function in R package "*minfi*" [14, 15].

4. Data Access

These Data are released through the UK Data Service. The End User Licence (EUL) version, SN7251, can be found here: https://beta.ukdataservice.ac.uk/datacatalogue/studies/study?id=7251.

5. Citation

If you use *Understanding Society* data you must cite every study that you use.

4.1 Citing the data

The citation for the data can be found at
https://www.understandingsociety.ac.uk/documentation/citation

All works which use or refer to these materials should acknowledge these sources by means of bibliographic citation. To ensure that such source attributions are captured for bibliographic indexes, citations must appear in footnotes or in the reference section of publications.

4.2 Citing this User Guide

When citing this User Guide, you can use the citation of this particular version quoted below.

Institute for Social and Economic Research (2025), Understanding Society: Waves 2-3 Nurse Health Assessment, 'Epigenetic ageing algorithms' derived from DNA methylation, 2010-2012, User Guide, Version 3, June 2025, Colchester: University of Essex.

6. Online resources and documentation

Information about Understanding Society health, biomarkers, genetics and epigenetic (DNA methylation) data can be found on the Understanding Society website at https://www.understandingsociety.ac.uk/documentation/health-assessment/. Specifically, you can find information on:

- How to access the data, user guides
- An interactive variable search facility
- Questionnaires
- Fieldwork documents
- Generic FAQs
- User guides. This section includes different user guides starting with the "Nurse Health Assessment User Guide" which describes the health and biomarker data collected by nurses. In addition to this user guide, this section includes the "Biomarker User Guide and Glossary", "Proteomics User Guide", "Polygenic Scores User Guide". This section also includes a "Biological data glossary".

7. References

[1] Bell CG, Lowe R, et.al. DNA methylation aging clocks: challenges and recommendations. *Genome Biology*, 2019, 20:249.

[2] Pidsley R, Wong CCY, et.al. A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics*, 2013, 14:293.

[3] Benzeval M, Devillas A, Kumari M and Lynn P. Understanding Society: The UK household longitudinal study – Biomarker user guide and glossary. 2014.

[4] Gorrie-Stone TJ, Smart MC, et.al. Bigmelon: tools for analysing large DNA methylation datasets, *Bioinformatics*, 2019, 35(6):981–986.

[5] Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biology*, 2013, 14:3156.

[6] Hannum G, Guinney J. Zhao L et.al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell*, 2013, 49(2):359-367.

[7] Levine ME, Lu At, Quach A, et.al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY)*, 2018, 10 (4): 573-591.

[8] Horvath S, Oshima J, Martin GM, et al. Epigenetic clock for skin and blood cells applied to Hutchinson Gilfor Progeria Syndrome and ex vivo studies. *Aging (Albany NY)*, 2018, 10(7): 1758-1775.

[9] Lin Q, Wagner W. Epigenetic aging signatures are coherently modified in cancer. *PLoS Genetics*, 2015, 11(6):e1005334.

[10] Weidner CI, Lin Q, Koch CM, et.al. Aging of blood can be tracked by DNA methylation changes at just three CpG sites. *Genome Biology*, 2014, R24.

[11] Lin Q, Weidner CI, Costa IG, et.al. DNA methylation levels at individual age-associated CpG sites can be indicative for life expectancy. *Aging (Albany NY)*, 2016, 8.2: 394-401.

[12] Belsky DW, Caspi A, Arseneault L, et.al. Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. E*Life*, 2020, May 5; 9:e54870.

[13] Belsky DW, Caspi A, Corcoran D, et.al. DunedinPACE, a DNA methylation biomarker of the pace of aging. *ELife*, 2022, Jan 14;11:e73420.

[14] Houseman EA, Accomando WP, Koestler DC, et.al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics*. 2012,13:86.

[15] Aryee MJ, Jaffe AE, Corrada-Bravo H, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics. 2014; 30(10):1363–1369

Supplementary -List of missing probes in ageing algorithms with the Illumina Infinium HumanMethylationEPIC array

Part 1: ***Horvath 2013 Missing Probes (up to 20 probes)***		
Probe ID	Chromosome	Gene Annotation
cg02654291	9	C9orf64
cg02972551	2	KDM3A;KDM3A
cg09785172	4	WFS1;WFS1;WFS1;WFS1
cg09869858	12	P11
cg13682722	14	C14orf102;C14orf102
cg14329157	2	WDR69
cg16494477	5	FGF18
cg17408647	7	C7orf44;C7orf44
cg19167673*	22	PDGFB
cg19273182	2	PAPOLG;PAPOLG
cg19945840	1	SDF4;SDF4;B3GALT6
cg27319898	7	ZNF804B;ZNF804B
cg27413543	4	SEC31A
cg04431054	5	PRRC1
cg05590257	17	PLD6
cg06117855	3	CLEC3B;CLEC3B
cg11025793	19	IER2;STX10
cg19046959	1	COL8A2
cg19569684	5	MGC29506
cg24471894	9	KIAA0020
cg27016307	19	HRC
* Only missing for sample measured at 2020 (sample size 2,840)		
Part 2*: Hannum Missing Probes (7 probes)*		
Probe ID	Chromosome	Gene Annotation
cg24079702	2	FHL2;FHL2;FHL2;FHL2
cg14361627	7	KLF14
cg22285878	7	KLF14
cg07927379	7	C7orf13;RNF32
cg18473521	12	HOXC4;HOXC4
cg09651136	15	PKM2;PKM2;PKM2
cg21139312	17	MSI2;MSI2
Part 3: ***PhenoAge Missing Probes (up to 2 probes)***		
Probe ID	Chromosome	Gene Annotation
cg08212685	5	ATG10
cg26665419**	6	C6orf203
** Only missing for samples measured at 2017 (sample size 1,174)		
Part 4: ***Lin Missing Probes (2 probes)***		
Probe ID	Chromosome	Gene Annotation
cg19046959	1	COL8A2
cg15379633	22	RAB36

Part 5: ***Belsky DunedinPoAm38 Missing Probes (1 probe)***		
Probe ID	Chromosome	Gene Annotation
cg06133392***	21	PCNT
*** Only missing for sample measured at 2020 (sample size 2,840)		
Part 6: ***Belsky DunedinPACE Missing Probes (1 probe)***		
Probe ID	Chromosome	Gene Annotation
cg11103390****	10	ECHDC3
**** Only missing for sample measured at 2020 (sample size 2,840)		