



OPEN

## A new blood based epigenetic age predictor for adolescents and young adults

Håvard Aanes<sup>1</sup>, Øyvind Bleka<sup>1</sup>, Pål Skage Dahlberg<sup>1</sup>, Kristina Totland Carm<sup>1</sup>, Terho Lehtimäki<sup>2</sup>, Olli Raitakari<sup>3,4,5</sup>, Mika Kähönen<sup>6</sup>, Mikko Hurme<sup>7,8</sup> & Veslemøy Rolseth<sup>1✉</sup>

Children have special rights for protection compared to adults in our society. However, more than 1/4 of children globally have no documentation of their date of birth. Hence, there is a pressing need to develop biological methods for chronological age prediction, robust to differences in genetics, psychosocial events and physical living conditions. At present, DNA methylation is the most promising biological biomarker applied for age assessment. The human genome contains around 28 million DNA methylation sites, many of which change with age. Several epigenetic clocks accurately predict chronological age using methylation levels at age associated CpG-sites. However, variation in DNA methylation increases with age, and there is no epigenetic clock specifically designed for adolescents and young adults. Here we present a novel age Predictor for Adolescents and Young Adults (PAYA), using 267 CpG methylation sites to assess the chronological age of adolescents and young adults. We compared different preprocessing approaches and investigated the effect on prediction performance of the epigenetic clock. We evaluated performance using an independent validation data set consisting of 18-year-old individuals, where we obtained a median absolute deviation of just below 0.7 years. This tool may be helpful in age assessment of adolescents and young adults. However, there is a need to investigate the robustness of the age predictor across geographical and disease populations as well as environmental effects.

Children are protected by a special subset of the human rights. However, as around 25% of the world's children do not hold a birth certificate<sup>1</sup> and therefore cannot document when they were born, age assessment is of great importance to secure children their human rights. Today, the most commonly applied age assessment methods in children and young adults include radiographs of teeth and skeleton, however, these are methods with large biological variation. In the recent decades, epigenetic clocks have emerged as a promising tool to predict both biological and chronological age.

Epigenetic age predictors are utilized to study both biological aging and in forensics. Environmental and psychological stressors affect epigenetic patterns<sup>2</sup>, and the development of epigenetic age predictors assessing biological age due to environmental influences or health issues are a growing field within aging research<sup>3-5</sup>. These predictors can, among other applications, be used to measure the effects of anti-aging interventions<sup>6</sup>. In contrast, in forensic applications, there is a need for predictors that are not affected by genetics, medical conditions or environmental variables (e.g. diet). Hence, epigenetic age predictors can be classified into two categories; one for forensic age assessment (chronological age), and another for health measurements (biological age). Forensic age estimation is of importance in a number of cases like unidentified bodies, suspects of crime, human trafficking, and in age assessment of asylum seekers with unknown age. These individuals may have been under immense physiological stress, and are likely to have experienced traumatic events during their lifetime. They may have

<sup>1</sup>Division of Laboratory Medicine, Department of Forensic Sciences, Oslo University Hospital, Nydalen, P.O. Box 4950, 0424 Oslo, Norway. <sup>2</sup>Department of Clinical Chemistry, Fimlab Laboratories and Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. <sup>3</sup>Centre for Population Health Research, University of Turku and Turku University Hospital, Turku, Finland. <sup>4</sup>Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland. <sup>5</sup>Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland. <sup>6</sup>Department of Clinical Physiology, Faculty of Medicine and Health Technology, Tampere University Hospital and Finnish Cardiovascular Research Center-Tampere, Tampere University, Tampere, Finland. <sup>7</sup>Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland. <sup>8</sup>Tampere University Hospital, Tampere, Finland. ✉email: vesrol@ous-hf.no

also experienced starvation and/or malnutrition. It is plausible that these stressors may influence the epigenetic markers in epigenetic clocks.

Epigenetics is the biology of genetic control without change of the genetic code. Often, chemical molecules are attached to the DNA to alter the way the DNA is read by specific intracellular proteins. Methylation, the attachment of methyl groups to specific sites, primarily cytosines followed by guanines are one of the most studied epigenetic mechanisms. These sites are referred to as CpG dinucleotides, and several repetitive CpGs are defined as a CpG island, often occurring near, or in the promoter region of genes. Among other functions, DNA methylation regulates transcription, most commonly inhibiting gene expression<sup>7</sup>.

DNA methylation plays an important role in gene regulation during development and aging, and consequently many sites are associated with age<sup>8</sup>. Numerous age predictors have been developed to estimate chronological age based on DNA methylation patterns<sup>9</sup>. These epigenetic clocks accurately predict chronological age, but are mainly derived from datasets consisting of individuals with a broad age range (i.e. 0–100 years). Several studies have shown increased variation in age associated methylation CpGs and predicted epigenetic age as chronological age increases<sup>10,11</sup>. Therefore, it is expected that epigenetic age predictions will be more precise for younger individuals.

As DNA methylation patterns change more rapidly in children and adolescents, the development of pediatric epigenetic clocks has been a recent focus, resulting in three available age predictors designed for a pediatric target group<sup>12–14</sup>. As previously mentioned, children are protected by specific rights, and therefore the age group of around 18 years of age is important to investigate in several forensic applications. This age group has, to our knowledge, not yet been specifically targeted by any epigenetic clock. Epigenetic analyses might be both time-consuming and expensive, but can be reduced by decreasing number of CpG sites analysed. Therefore, age predictors developed for forensic use, are typically based on few CpG sites<sup>15,16</sup>. However, such predictors have lower accuracy than models with several hundred methylation sites, and better prediction models are warranted to assess if an individual is a child or an adult.

DNA methylation levels are commonly analysed using Illumina DNA methylation arrays. When it was introduced in 2007, the Illumina array covered 27,000 CpG sites, which was later upgraded to 450,000 sites in 2010 (450 K), and again to 850,000 sites in the latest EPIC array (2015). These arrays have high accuracy and precision<sup>17</sup>.

The analysis of DNA methylation microarrays is complex and there are numerous methods available aiming to improve data quality. These methods can be broadly divided into preprocessing, data normalization and batch correction, and each one can influence the downstream results<sup>18</sup>. The first step, preprocessing, includes image processing, calculation of methylation levels, quality control and filtering of probes and chips. The second step, normalization, intends to remove technical variation between and/or within chips. Normal-exponential out-of-band (Noob) is a commonly used method for background correction, which also includes dye bias correction<sup>19</sup>. Finally, batch effects; e.g. samples analysed on different days, position on the chip or efficiency of the bisulphite conversion, may systematically differ. ComBat<sup>20</sup> is a widely conducted batch correction method. However as for most batch correction methods, warnings of its use leading to masking of biological differences are frequent<sup>21,22</sup>.

Machine learning methods are commonly applied to make age predictors using DNA methylation data. By far, the most popular method is elastic net (EN)<sup>23</sup>. EN is a regularized regression method that simultaneously selects the best set of methylation sites, and shrinks the coefficients by applying a penalty parameter. During the training of the model, the dependent variable is age, and the methylation sites are the independent variables. The trained model is then used to predict the age of new individuals. Predictive models are prone to overfitting, that is, they are parameterized well to the training data, but can perform poorly when predicting new data. This is due to variation in the training data included in the model, but not truly related to the outcome (e.g. age in our case)<sup>24</sup>. However, the result of the EN analysis is a small number of sites (relative to what you start with) where overfitting is reduced to some extent, and enhanced prediction performance for new data is achieved.

In the present study, we have developed the PAYA age predictor; a blood-based age predictor for adolescents and young adults between 12 and 25 years old, using DNA methylation levels at 267 CpG sites, aimed to assess chronological age without interference from environmental and disease conditions. This predictor might be included as part of a data driven process for assessing the chronological age of individuals in this age group.

## Methods

**Study participants.** We identified relevant studies and downloaded data from GEO (Supplementary file 1—Supplementary Table S1). Inclusion criteria were studies using 450 K array data, generated from blood samples of individuals in the age range 10–60 years. Information on age and access to IDAT files (raw data) were also necessary. In addition, we included 450 K array data from the Young Finns study (YFS). The total training dataset consisted of 2316 samples from 1013 males and 1303 females. To test the model, we used 450 K array data from 920 18-year old individuals from the E-risk study<sup>25</sup>, hereby termed the test dataset. This is a birth cohort of twins born in 1994–5 in the United Kingdom. We randomly included one twin from each pair in our test group, in total 454 males and 466 females.

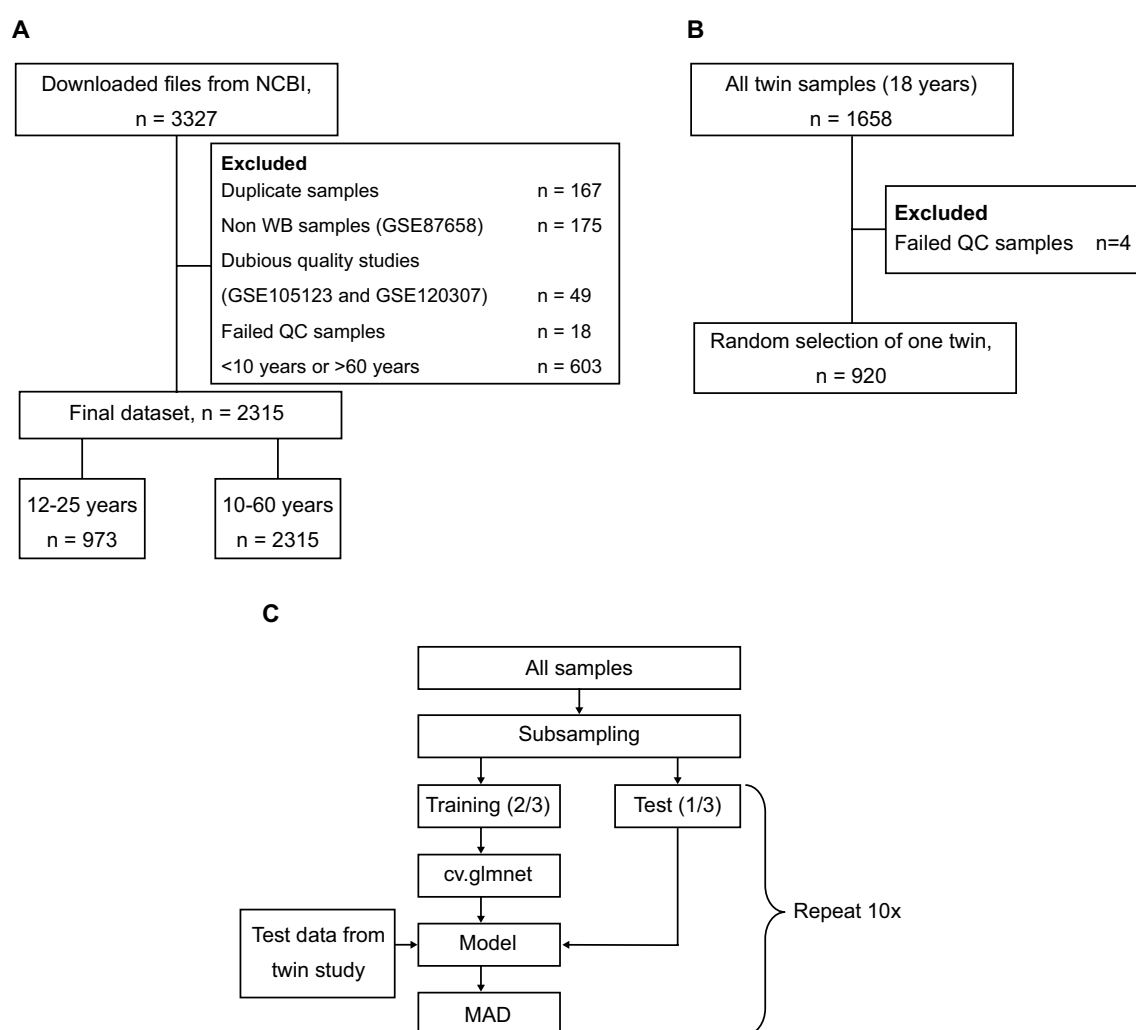
**Methylation analysis pipeline.** DNA methylation data (IDAT files) was imported into R with the “read.metharray.exp” command in the Bioconductor package Minfi<sup>26</sup>. The methylation levels were calculated as beta-values:  $\text{Methylated}/(\text{Methylated} + \text{Unmethylated} + 100)$ . During import we estimated cell counts using the “estimateCellCounts” function in Minfi<sup>27</sup> and normalised data using “preprocessNoob” (includes background correction and dye-bias correction).

We designed a rigorous quality control pipeline to detect samples of low quality (see Supplementary file 1—Supplementary methods), and samples were removed based on manual inspection of the quality control summary (see Supplementary file 1 for an overview of excluded sites and reason for exclusion). Based on this quality

control, two studies were left out (Supplementary file 1—Supplementary methods). The final training dataset consisted of 2316 samples (age span 10–60 years old, 1013 males and 1303 females see Fig. 1a and Supplementary figure S1 for details). From the test dataset, we excluded eight samples due to poor quality, and included only one randomly selected individual from each twin pair (Fig. 1b). Probes associated with SNPs, and sites reported to have cross-reactivity, as well as probes not found in the newer Infinium EPIC chip, were removed (See [https://github.com/sirselim/illumina450k\\_filtering](https://github.com/sirselim/illumina450k_filtering) for details). A total of 401,484 sites were removed after this filtering.

**Evaluation of prediction performance.** To evaluate the impact of different processing pipelines we constructed a function in R to run elastic net cross validation repeatedly: We used the `cv.glmnet` function from GLMnet R-package (v4.1–3) with default loss metric,  $\alpha = 0.5$ , and the minimum lambda value chosen (if not otherwise stated). The training data was sampled into a random training subset (2/3 of the samples), and used in the cross validation procedure, and a test set (1/3 of the samples) were used to predict age. In addition, we evaluated the performance of the built model from the cross validation using the twin test dataset, as these data were not part of the cross-validation procedure (Fig. 1c). Hence, we had two test data in this function. As our measure of performance, we used the median absolute deviation (MAD) measure (years between predicted and chronological age). The training and prediction was repeated ten times to obtain a distribution of MAD values. In the final model fitting (i.e. the final predictor) all samples, except the independent twin test dataset, were used as training data.

**Ethical considerations.** All methods in the present study were carried out in accordance with the relevant guidelines and regulations. The Young Finns study was approved by the ethical committee of the Hospital



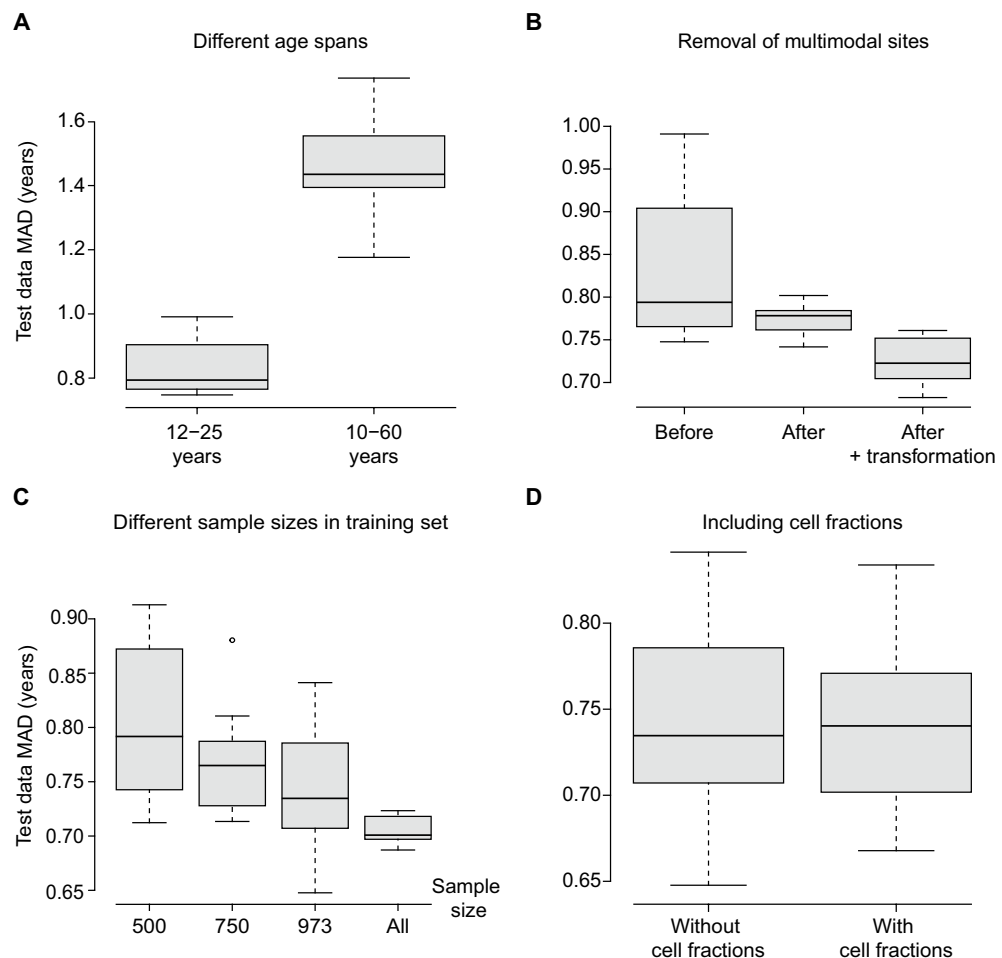
**Figure 1.** Sample retrieval and filtering. (A) Selection and quality control of the training data. (B) Selection and quality control of the independent test dataset. (C) The performance evaluation procedure. Using the training data, we selected 2/3 of the samples to fit an age prediction model using elastic net and the cross validation method. The remaining 1/3 of the samples were used to test the data, and the test dataset was also tested using the generated model. As a measure of prediction performance we used the median absolute deviation (MAD) in both cases.

District of Southwest Finland on 20 June 2017 (ETMK:68/1801/2017) and Regional Ethics Committee of the Expert Responsibility area of Tampere University Hospital, Helsinki University Hospital Ethical Committee of Medicine, The Research Ethics Committee of the Northern Savo Hospital District and Ethics Committee of the Northern Ostrobothnia Hospital District. The study protocol of each study phase corresponded to the WHO proposal. All participants gave their written informed consent, and the studies were conducted in accordance with the Declaration of Helsinki. At prior follow-ups of the Young Finns Study, informed consent of every participant under the age of 18 was obtained from a parent and/or legal guardian. The remaining datasets used in the present study were collected from previously published sources (Accession numbers in Supplementary table S1), and an approval by an ethics committee in the use of these data was therefore not necessary. In all included studies, authors state that informed consent for each participant was obtained.

## Results

**Age span of the training data and age transformation.** We tested the impact of building models with a narrow (12–25 years,  $n=973$  samples) or broad age span (10–60 years,  $n=2316$  samples). There was a distinct performance difference between the two approaches, where using the more narrow dataset improved performance (Fig. 2a), despite having less than half the number of samples. The age predictor trained on a wider age span was more accurate when we added “Horvath’s transformation”<sup>78</sup> (described in Supplementary methods and shown in Supplementary Fig. S2a), while the 12–25 years predictions were not substantially affected (Supplementary Fig. S2b). Nonetheless, we chose to use the transformation for further analysis. The intended use of this model is in an adolescent population; we therefore chose to use the approach trained on the narrow age span.

**Removal of multimodal sites.** We observed that some DNA methylation sites were included in the models despite not correlating with age, but instead having the appearance of outliers (Supplementary file 2). Consequently, we made a function to detect multimodal sites, and subsequently removed them (implemented in `normtools`, see Supplementary methods in Supplementary file 1). This reduced the number of sites to 393,821.



**Figure 2.** Impact of different methodological factors on the performance of the model. (A) Impact of using different age spans. (B) Removal of multimodal sites. (C) Impact of different sample sizes. (D) Including cell fractions in the model.

While this intervention did not influence the median of the MAD much, it did reduce the variability considerably (Fig. 2b), and we decided to use this filtered dataset for further explorations.

**Sample size.** The number of samples have been shown to affect the performance of age prediction models<sup>28,29</sup>. Down-sampling of our dataset to 500 and 750 randomly chosen samples, led to a gradual reduction in performance (Fig. 2c). We also measured performance without subsampling, i.e. using all samples for each model building. This resulted in further improvement of performance, with MAD around 0.70 years (min–max 0.69–0.72).

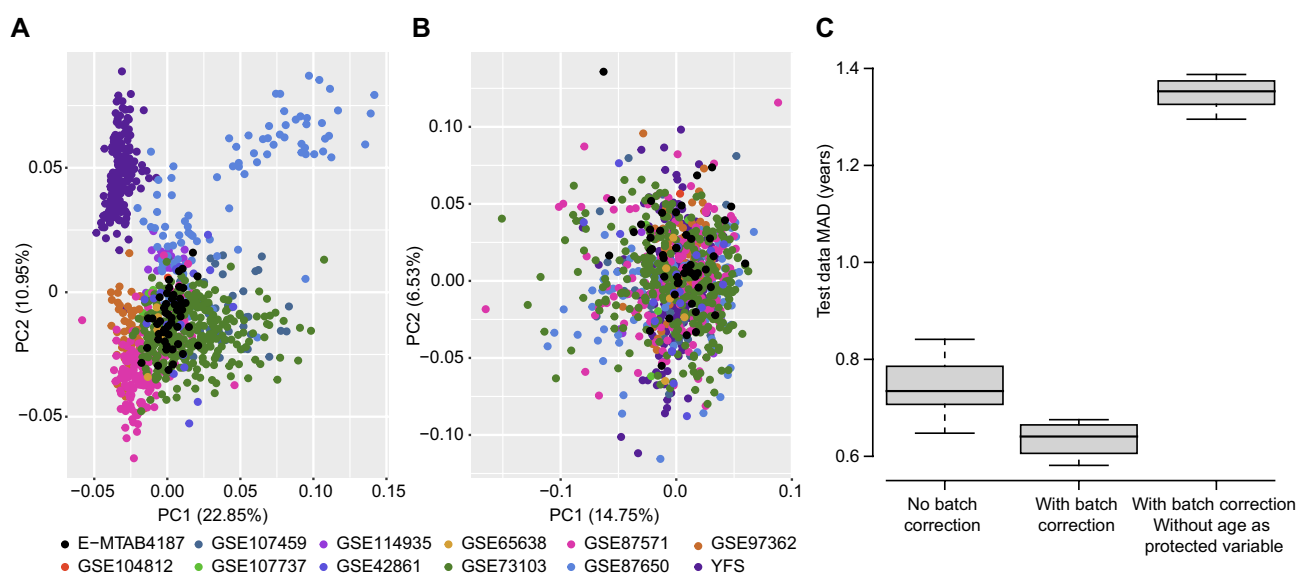
**Using cell counts in the predictions.** It has been suggested that cell type composition might be an important variable in explaining DNA methylation patterns<sup>30</sup>. To test this we compared model building with and without predicted cell type fractions<sup>31</sup>. There were no improvement of age predictions for the test dataset when including cell type information, and therefore no cell type fractions were included in the model (Fig. 2d).

**Batch correction.** We applied principal component analysis (PCA) to uncover possible batch effects in our included data. PCA showed clustering of studies, indicative of batch effects (Fig. 3a). To overcome these batch effects, we applied ComBat<sup>20,32</sup>. We first used the default method (parametric version i.e. all studies become relative to each-other), and included age as a protected variable. This resulted in large differences between the training- and the independent test dataset (data not shown). However, after switching to “reference study mode” (adjusting training- and test data using the YFS study), the two datasets became comparable. After batch correction, the training data did no longer cluster by study (Fig. 3b), and predictions improved for the test data, compared to using no batch corrections (Fig. 3c). However, in a forensic application, the age of individuals tested will be unknown, hence we cannot use the chronological age as a “protected variable”. Without this, the use of batch correction did not improve predictions (Fig. 3c), on the contrary, the performance deteriorated considerably.

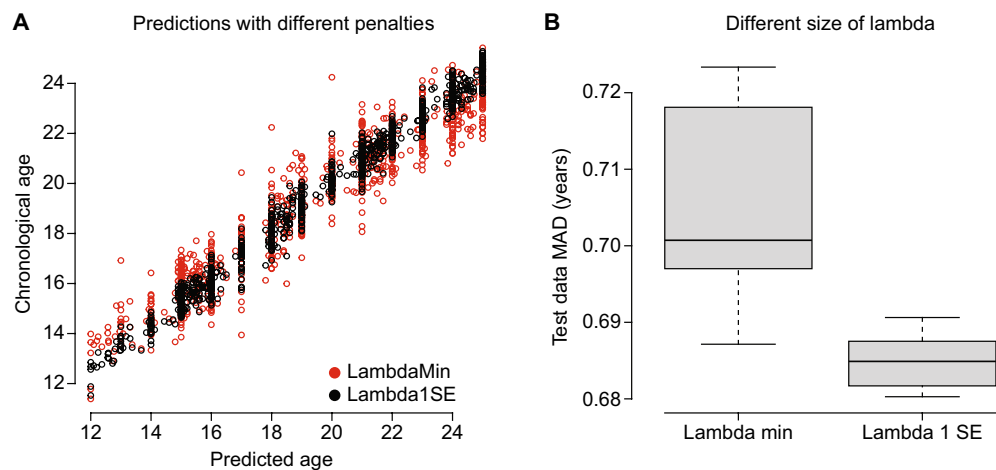
To by-pass the issue of not knowing chronological age, we instead tried to use the predicted age from the built model as the “protected variable” in ComBat. However, this strategy did not outperform using Noob normalization only (data not shown).

**Building the final model and marginal inspections.** After testing various preprocessing and normalization options (see methods and Supplementary Fig. S3), we opted for the following approach to build PAYA; we normalised data with Noob (dye bias correction included), before we removed samples and probes of low quality. We then removed CpG sites with multinomial modes and transformed the age variable using Horvath’s transformation of age (described in Supplementary methods). We trained the model on a narrow age span, according to the intended application of the method.

In EN, cross validation is used to calibrate the lambda penalty values. The lambda is typically chosen such that it returns the smallest cross validation error (i.e. “lambda.min”). When considering only the training data, smaller MAD were observed using “lambda.min” over “lambda.1se” (MAD 0.33 vs 0.73 years) (Fig. 4a). However, using “lambda.1se” returned slightly, but consistently smaller MAD for the test dataset (Fig. 4b): Also, we obtained using only 267 vs 660 CpG sites with the larger lambda.



**Figure 3.** Identification and removal of batch effects. (A) Principal component analysis prior to batch effect removal. (B) Principal component analysis after batch effect removal. (C) Prediction performance in the test dataset before and after batch effect removal, and without age as a “protected variable”.



**Figure 4.** Choosing lambda. (A) Scatter plot of predicted and chronological age using different penalties (minimum lambda in black, one standard deviation lambda in red). (B) Prediction performance on independent test dataset using the same lambdas.

Methylation level by age in the training dataset were plotted for all sites included in the model (Supplementary file 2). Correlations (Spearman) between degree of methylation and age, together with their associated coefficients, can be found in Supplementary file 3. Despite removing the multimodal sites, we observe some sites included in the final model that appear multimodal (e.g. cg21015022). In addition, we find sites that do not correlate well with age (e.g. cg03846689), and contained outlier values.

All sites in the age predictor were annotated with associated gene names (Supplementary file 4). Functional annotation and disease association analysis in Metascape<sup>33</sup> was conducted. The analysis revealed no enrichment in the functional annotation categories, however, some disease associated terms were enriched (mental disorders, gait abnormality and smoking), with borderline significant q-values (Supplementary file 5).

We observed that most predictions were within  $\pm 1$  year (of 18.5) (600/920, 65%), but there were also some predictions deviating  $> 2$  years (45,  $\sim 5\%$ ), and five observations deviating  $> 3$  years. Of those deviating  $> 3$  years, three were overestimated (predicted ages 21.6, 22.1 and 22.2) and two underestimated (15.4 and 15.5), with the most severe error deviating 3.7 years. See Supplementary Fig. S4 for histogram of the observed errors. We found no effect of gender on the age-prediction. Using our test data (18 year olds), mean age prediction was 18.2 years for both genders.

**Comparison with other age predictors and studies.** We compared the sites in our model with sites in other age predictor models utilizing DNA methylation (see Table 1 for details). Most overlap was observed between the Zhang epigenetic clock<sup>29</sup> and the cABEC clock<sup>28</sup>, with 47 and 41 sites overlapping with our model, respectively (Table 1). Horvath's "skin and blood predictor" shared 26 sites with our predictor. Little overlap was

Epigenetic clock	Publication	Number of CpG sites	Age span	Training, number of samples	Tissue	Overlapping CpG sites with PAYA	%
PAYA	Aanes et al., 2023	267	12–25	2315	Blood	–	–
Horvath	Horvath et al., 2013 <sup>8</sup>	353	0–100	3931	27 different	4	1.5
Hannum	Hannum et al., 2013 <sup>37</sup>	71	19–101	656	Blood	14	5.2
PhenoAge	Levine et al., 2018 <sup>3</sup>	513	18–100	926	Blood	6	2.2
PedBE	McEwen et al., 2020 <sup>12</sup>	94	0–20	1032	Buccal	5	1.9
Skin and Blood	Horvath et al., 2018 <sup>36</sup>	391	0–94	896	Blood, more*	26	9.7
DNAmTL	Lu et al., 2019 <sup>34</sup>	140	21–100	2256	Blood	4	1.5
Wu	Wu et al., 2019 <sup>13</sup>	111	0–18	716	Blood	1	0.4
Zhang	Zhang et al., 2019 <sup>29</sup>	514	2–104	13,661	Blood, saliva	47	17.6
Li	Li et al., 2018 <sup>14</sup>	83	6–17	90	Blood	2	0.7
cABEC	Lee et al., 2020 <sup>28</sup>	1892	19–88	2227	Blood	41	15.4

**Table 1.** Overview of epigenetic age predictors and the overlap in CpG sites compared to PAYA (Predictor of Adolescents and Young Adults). \*Human fibroblasts, keratinocytes, buccal cells, endothelial cells and skin.

found with other blood based paediatric age predictors<sup>13,14</sup>. As expected, little overlap with the telomere age predictor was recognised<sup>34</sup>.

Due to lack of information on the preprocessing steps leading up to beta values in other epigenetic age predictors, direct comparisons are difficult to conduct. However, we ran our test dataset (age 18) through different epigenetic clocks, using the R-package methylClock<sup>35</sup>. Results from four of these clocks (Horvath, Skin and Blood, Hannum and the Elastic Net clock by Zhang et al.<sup>8,29,36,37</sup>) showed that PAYA predicted chronological age with the highest precision and accuracy (Supplementary Fig. S5).

**Assessment of the impact of stress on sites in the age predictor.** To evaluate if our age predictor could be affected by childhood trauma and possible physiological stress, we generated a DNA methylation site list from a review of DNA methylation effects from childhood trauma<sup>38</sup> (Supplementary file 6). Only one CpG site from this list was part of our age predictor, cg07012999. This site has a small positive coefficient of 0.1, and therefore a marginal effect in the age prediction. Two *FKBP5* associated CpG sites, cg20813374 and cg00130530, linked to age and stress<sup>39</sup>, were not found in our model.

To our knowledge, no epigenome wide studies to investigate the impact of starvation and/or malnutrition at the epigenetic level for adolescents or young adults have been conducted. Two studies have, however, investigated locus specific effects on the *IGF2* gene<sup>40,41</sup>. This gene is not present in our cohort of genes associated with the sites of the age predictor.

## Discussion

In this study, we present a novel blood-based epigenetic age predictor, PAYA, based on data from adolescents and young adults between 12 and 25 years old. The predictor includes 267 CpG sites and showed a high degree of accuracy in prediction of chronological age in an independent test set of 18-year-old individuals with an estimate of MAD value below 0.7 years. This is noticeably better than targeted approaches with fewer CpG sites, which typically report MAD values between 5 and 9 years<sup>42</sup>.

Although comparisons with other epigenome-wide age predictors are difficult due to different data and normalization used for training, we find that PAYA perform better in the independent test data compared to other epigenetic clocks (Supplementary Fig. S5). We observed that reduction of the age span of the training data caused selection of other CpG sites, and increased prediction performance considerably compared to a wider age span (Fig. 2A). This is in concordance with previously reported results<sup>12</sup>.

As mentioned, the age range of the included samples might affect the included CpG sites in the predictor. To our surprise, we observed little overlap of CpG sites with other pediatric clocks. Pediatric clocks have been evaluated by others, and were outperformed by the Skin and Blood clock from Horvath et al.<sup>43</sup>. In line with this, our model shares more sites with the Skin and Blood clock compared to two pediatric clocks (Table 1). We observe that PAYA assess age more accurate in the independent test population than the Skin and Blood clock (Supplementary Fig. S5), probably due to the targeted age span on which it is trained.

Our model is trained and tested using Infinium HumanMethylation450 BeadChip data, but since new data will originate from the EPIC arrays, we have selected only EPIC compatible CpG sites to build our model. Despite the much larger number of sites included in the EPIC-array, it does not appear that epigenetic age predictors derived from EPIC data achieve higher accuracy than predictors based on the 450 K array<sup>28,44</sup>.

It has been claimed by others that increasing the number of samples will make the model more accurate<sup>28,29</sup>. This was confirmed in our study, as down-sampling of our own data resulted in lower accuracy of the age predictor (Fig. 2C). Notably, we did not observe a horizontal asymptote of the MAD value when all data was included in the training. Hence, the accuracy of the present model might be improved if more IDAT files were available.

The two most important tunable parameters in the EN analysis are the alpha and lambda values. The former decides the amount of Lasso versus Ridge regression, while lambda is the amount of penalty used<sup>45</sup>. We did initial testing of alpha values, and opted for a value of 0.5, similar to most existing models<sup>8,28,29,36</sup>. Lambda was chosen using k-fold cross-validation (with k = 10), and prediction performance on the independent test dataset. Interestingly, we achieved better performance when we chose the largest lambda within one standard error of the minimum prediction error providing less included CpG sites (from > 600 to < 300). We validated the trained age predictor utilizing 920 individuals (18 year olds), from an independent dataset. It is crucial to test the age predictors on an independent dataset since there can be study-specific factors that are explained by the model, but not present in independent data<sup>46</sup>. We observed similar performance between the test data and the independent test dataset, indicating that overfitting has been avoided.

It has been suggested that environmental effects are particularly important in epigenetic variation<sup>47</sup>. This confounding variable, as well as a number of diseases and conditions remains an important area for future research<sup>47,48</sup>. Recently, Mayer and colleagues revealed dependency between epigenetic age markers and growth disorders<sup>49</sup>. The intended use of PAYA is to predict chronological age not influenced by genetic and environmental effects. We tried to overcome this to some degree by including studies from different parts of the world, and only excluded conditions that are readily observable, meaning that not only healthy individuals are part of the data. The sites identified by EN may therefore be independent of some genetic and environmental effects, because if they were not, they would perhaps not explain the observed variation in chronological age well. PAYA might therefore, to some extent capture such effects. However, to fully assess the robustness of the current age predictor across populations with different ethnicities and environmental conditions, we have initiated a study where we sample individuals from different regions of the world, and evenly over the age span between 12 and 25 years of age.

The data used to build our model was from blood samples, and therefore contain several different cell types. The cell type composition of the blood might influence the obtained DNA methylation profiles<sup>50</sup>. It has been

suggested that the changes in cell type composition occurring with age, at least in part, may explain the age associated changes of DNA methylation<sup>51</sup>. However, we could not identify any effect by including cell type fractions in the age predictors (Fig. 2D). Other cell type prediction methods could perhaps yield better explanatory variables in the model<sup>52</sup>, but this remains to be tested. There is also a need to evaluate the actual composition of the blood and not only the predicted composition.

Despite our initial tests of the age predictor, a more thorough evaluation of performance is needed, with a full panel of performance metrics<sup>53</sup>. Application of the age predictor in forensic work will require thorough quality control (Ref Supplementary file 1) of new samples. It will also be of interest to test the PAYA age predictor in populations outside the intended age span to investigate the accuracy outside the age range of training.

To conclude, through thorough evaluation and selection of different analytical options, we have developed PAYA, the first blood based age predictor developed specifically for adolescents and young adults. We observe that PAYA outperforms existing epigenetic age predictors, making it eligible for application in forensic age-assessment, either alone or in combination with existing radiographic methods. Future studies is required to reveal if PAYA is robust when applied to diverse populations with different ethnicities and environmental or genetic effects.

## Data availability

The datasets used and/or analysed in this article are freely available datasets from each of the included studies (Supplementary table S1), except for the dataset obtained from the Cardiovascular Risk In Young Finns study (YFS) after submission and approval of our study plan by the YFS coordinators. The YFS dataset comprises health related participant data and their use is therefore restricted under the regulations on professional secrecy (Act on the Openness of Government Activities, 612/1999) and on sensitive personal data (Personal Data Act, 523/1999, implementing the EU data protection directive 95/46/EC). Due to these legal restrictions, the data from this study can not be stored in public repositories or otherwise made publicly available. However, data access may be permitted on a case by case basis upon request only. Data sharing outside the group is done in collaboration with YFS group and requires a data-sharing agreement. Investigators can submit an expression of interest to the chairman of the publication committee (Prof Mika Kähönen, Tampere University, Finland).

Received: 28 October 2022; Accepted: 3 February 2023

Published online: 09 February 2023

## References

1. United Nations Children's Fund, Birth Registration for Every Child by 2030: Are we on track?, UNICEF, New York, 2019.
2. Oblak, L., van der Zaag, J., Higgins-Chen, A. T., Levine, M. E. & Boks, M. P. A systematic review of biological, social and environmental factors associated with epigenetic clock acceleration. *Ageing Res. Rev.* **69**, 101348. <https://doi.org/10.1016/j.arr.2021.101348> (2021).
3. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591. <https://doi.org/10.18632/aging.101414> (2018).
4. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany N.Y.)* **11**, 303–327. <https://doi.org/10.18632/aging.101684> (2019).
5. Belsky, D. W. *et al.* DunedinPACE, a DNA methylation biomarker of the pace of aging. *Elife* **11**, 259. <https://doi.org/10.7554/eLife.73420> (2022).
6. Fahy, G. M. *et al.* Reversal of epigenetic aging and immunosenescent trends in humans. *Aging Cell* **18**, 028. <https://doi.org/10.1111/acel.13028> (2019).
7. Moore, L. D., Le, T. & Fan, G. DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38. <https://doi.org/10.1038/npp.2012.112> (2013).
8. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, 1151. <https://doi.org/10.1186/gb-2013-14-10-r115> (2013).
9. Field, A. E. *et al.* DNA methylation clocks in aging: Categories, causes, and consequences. *Mol. Cell* **71**, 882–895. <https://doi.org/10.1016/j.molcel.2018.08.008> (2018).
10. van Dongen, J. *et al.* Genetic and environmental influences interact with age and sex in shaping the human methylome. *Nat. Commun.* **7**, 11115. <https://doi.org/10.1038/ncomms11115> (2016).
11. Sliker, R. C. *et al.* Age-related accrual of methylomic variability is linked to fundamental ageing mechanisms. *Genome Biol.* **17**, 191. <https://doi.org/10.1186/s13059-016-1053-6> (2016).
12. McEwen, L. M. *et al.* The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 23329–23335. <https://doi.org/10.1073/pnas.1820843116> (2020).
13. Wu, X. *et al.* DNA methylation profile is a quantitative measure of biological aging in children. *Aging (Albany N.Y.)* **11**, 10031–10051. <https://doi.org/10.18632/aging.102399> (2019).
14. Li, C. *et al.* Age prediction of children and adolescents aged 6–17 years: An epigenome-wide analysis of DNA methylation. *Aging (Albany N.Y.)* **10**, 1015–1026. <https://doi.org/10.18632/aging.101445> (2018).
15. Freire-Aradas, A. *et al.* A common epigenetic clock from childhood to old age. *Forens. Sci. Int. Genet.* **60**, 102743. <https://doi.org/10.1016/j.fsigen.2022.102743> (2022).
16. Koch, C. M. & Wagner, W. Epigenetic-aging-signature to determine age in different tissues. *Aging* **3**, 1018–1027. <https://doi.org/10.18632/aging.100395> (2011).
17. Pidsley, R. *et al.* Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling. *Genome Biol.* **17**, 208. <https://doi.org/10.1186/s13059-016-1066-1> (2016).
18. Lussier, A. A. *et al.* Updates to data versions and analytic methods influence the reproducibility of results from epigenome-wide association studies. *Epigenetics* **1–16**, 2022. <https://doi.org/10.1080/15592294.2022.2028072> (2022).
19. Fortin, J. P., Triche, T. J. Jr. & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinform. (Oxf. Engl.)* **33**, 558–560. <https://doi.org/10.1093/bioinformatics/btw691> (2017).
20. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostat. (Oxf. Engl.)* **8**, 118–127. <https://doi.org/10.1093/biostatistics/kxj037> (2007).
21. Zindler, T., Frieling, H., Neyazi, A., Bleich, S. & Friedel, E. Simulating ComBat: How batch correction can lead to the systematic introduction of false positive results in DNA methylation microarray studies. *BMC Bioinform.s* **21**, 271. <https://doi.org/10.1186/s12859-020-03559-6> (2020).



22. Nygaard, V., Rødland, E. A. & Hovig, E. Methods that remove batch effects while retaining group differences may lead to exaggerated confidence in downstream analyses. *Biostat. (Oxf., Engl.)* **17**, 29–39. <https://doi.org/10.1093/biostatistics/kxv027> (2016).
23. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **67**, 301–320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x> (2005).
24. Simon, R., Radmacher, M. D., Dobbin, K. & McShane, L. M. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J. Natl. Cancer Inst.* **95**, 14–18. <https://doi.org/10.1093/jnci/95.1.14> (2003).
25. Hannon, E. *et al.* Characterizing genetic and environmental influences on variable DNA methylation using monozygotic and dizygotic twins. *PLoS Genet.* **14**, e1007544. <https://doi.org/10.1371/journal.pgen.1007544> (2018).
26. Aryee, M. J. *et al.* Minfi: A flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinform. (Oxf., Engl.)* **30**, 1363–1369. <https://doi.org/10.1093/bioinformatics/btu049> (2014).
27. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinform.* **13**, 86. <https://doi.org/10.1186/1471-2105-13-86> (2012).
28. Lee, Y. *et al.* Blood-based epigenetic estimators of chronological age in human adults using DNA methylation data from the Illumina MethylationEPIC array. *BMC Genom.* **21**, 747. <https://doi.org/10.1186/s12864-020-07168-8> (2020).
29. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* **11**, 54. <https://doi.org/10.1186/s13073-019-0667-1> (2019).
30. Teschendorff, A. E. & Zheng, S. C. Cell-type deconvolution in epigenome-wide association studies: A review and recommendations. *Epigenomics* **9**, 757–768. <https://doi.org/10.2217/epi-2016-0153> (2017).
31. Gervin, K. *et al.* Cell type specific DNA methylation in cord blood: A 450K-reference data set and cell count-based validation of estimated cell type composition. *Epigenetics* **11**, 690–698. <https://doi.org/10.1080/15592294.2016.1214782> (2016).
32. Jeffrey, T. L. *et al.* Storey and Leonardo Collado & Torres. *sva: Surrogate Variable Analysis. R package version 3.40.0. R Package* (2021).
33. Zhou, Y. *et al.* Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. *Nat. Commun.* **10**, 1523. <https://doi.org/10.1038/s41467-019-09234-6> (2019).
34. Lu, A. T. *et al.* DNA methylation-based estimator of telomere length. *Aging* **11**, 5895–5923. <https://doi.org/10.18632/aging.102173> (2019).
35. Pelegí-Sisó, D., de Prado, P., Ronkainen, J., Bustamante, M. & González, J. R. methylclock: A bioconductor package to estimate DNA methylation age. *Bioinform. (Oxf., Engl.)* **37**, 1759–1760. <https://doi.org/10.1093/bioinformatics/btaa825> (2021).
36. Horvath, S. *et al.* Epigenetic clock for skin and blood cells applied to Hutchinson Gilford Progeria Syndrome and ex vivo studies. *Aging* **10**, 1758–1775. <https://doi.org/10.18632/aging.101508> (2018).
37. Hannum, G. *et al.* Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367. <https://doi.org/10.1016/j.molcel.2012.10.016> (2013).
38. Womersley, J. S. *et al.* Childhood trauma, the stress response and metabolic syndrome: A focus on DNA methylation. *Eur. J. Neurosci.* **55**, 2253–2296. <https://doi.org/10.1111/ejn.15370> (2022).
39. Zannas, A. S. *et al.* Epigenetic upregulation of FKBP5 by aging and stress contributes to NF-κB-driven inflammation and cardiovascular risk. *Proc. Natl. Acad. Sci. USA* **116**, 11370–11379. <https://doi.org/10.1073/pnas.1816847116> (2019).
40. Shen, L. *et al.* Early-life exposure to severe famine is associated with higher methylation level in the IGF2 gene and higher total cholesterol in late adulthood: The Genomic Research of the Chinese Famine (GRECF) study. *Clin. Epigenet.* **11**, 88. <https://doi.org/10.1186/s13148-019-0676-3> (2019).
41. Heijmans, B. T. *et al.* Persistent epigenetic differences associated with prenatal exposure to famine in humans. *Proc. Natl. Acad. Sci.* **105**, 17046–17049. <https://doi.org/10.1073/pnas.0806560105> (2008).
42. Parson, W. Age estimation with DNA: From forensic DNA fingerprinting to forensic (Epi)genomics: A mini-review. *Gerontology* **64**, 326–332. <https://doi.org/10.1159/000486239> (2018).
43. Kling, T., Wenger, A. & Carén, H. DNA methylation-based age estimation in pediatric healthy tissues and brain tumors. *Aging (Albany NY)* **12**, 21037–21056. <https://doi.org/10.18632/aging.202145> (2020).
44. Alsaleh, H. & Hadrill, P. R. Identifying blood-specific age-related DNA methylation markers on the Illumina MethylationEPIC® BeadChip. *Forensic Sci. Int.* **303**, 109944. <https://doi.org/10.1016/j.forsciint.2019.109944> (2019).
45. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* 2nd edn. (Springer, 2009).
46. Teschendorff, A. E. Avoiding common pitfalls in machine learning omic data science. *Nat. Mater.* **18**, 422–427. <https://doi.org/10.1038/s41563-018-0241-z> (2019).
47. Li, S. *et al.* Genetic and environmental causes of variation in epigenetic aging across the lifespan. *Clin. Epigenet.* **12**, 158. <https://doi.org/10.1186/s13148-020-00950-1> (2020).
48. Koop, B. E. *et al.* Epigenetic clocks may come out of rhythm—implications for the estimation of chronological age in forensic casework. *Int. J. Legal Med.* **134**, 2215–2228. <https://doi.org/10.1007/s00414-020-02375-0> (2020).
49. Mayer, F. *et al.* Altered DNA methylation at age-associated CpG sites in children with growth disorders: Impact on age estimation?. *Int. J. Legal Med.* **136**, 987–996. <https://doi.org/10.1007/s00414-022-02826-w> (2022).
50. Liu, Y. *et al.* Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat. Biotechnol.* **31**, 142–147. <https://doi.org/10.1038/nbt.2487> (2013).
51. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **19**, 371–384. <https://doi.org/10.1038/s41576-018-0004-3> (2018).
52. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* **19**, 129–147. <https://doi.org/10.1038/nrg.2017.86> (2018).
53. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genom.* **13**, S2. <https://doi.org/10.1186/1471-2164-13-S4-S2> (2012).

## Acknowledgements

We want to thank all the researchers who generously made their data public available. The Young Finns Study has been financially supported by the Academy of Finland: grants 322098, 286284, 134309 (Eye), 126925, 121584, 124282, 255381, 256474, 283115, 319060, 320297, 314389, 338395, 330809, 104821, 129378 (Salve), 117797 (Gendi), and 141071 (Skidi); the Social Insurance Institution of Finland; Competitive State Research Financing of the Expert Responsibility area of Kuopio, Tampere and Turku University Hospitals (grant X51001); Juho Vainio Foundation; Paavo Nurmi Foundation; Finnish Foundation for Cardiovascular Research; Finnish Cultural Foundation; The Sigrid Juselius Foundation; Tampere Tuberculosis Foundation; Emil Aaltonen Foundation; Yrjö Jahansson Foundation; Signe and Ane Gyllenberg Foundation; Diabetes Research Foundation of Finnish Diabetes Association; EU Horizon 2020 (grant 755320 for TAXINOMISIS and grant 848146 for To Aition); European Research Council (grant 742927 for MULTIEPIGEN project); Tampere University Hospital Supporting Foundation, Finnish Society of Clinical Chemistry and the Cancer Foundation Finland. TL, OR, MK and MH were funded by the Young Finns Study.

### Author contributions

H.A.: methodology, validation, formal analysis, investigation, writing—original draft, visualization; Ø.B.: methodology, software, validation, formal analysis, investigation, writing—original draft; P.S.D.: conceptualization, methodology, resources, writing—review & editing; K.T.C.: writing—review & editing, visualization; T.L.: resources; O.R.: resources, M.K.: resources; M.H.: resources; V.R.: conceptualization, project administration, methodology, investigation, resources, writing—original draft. All authors reviewed the final manuscript.

### Competing interests

The Norwegian Directorate of Immigration finances the academic work concerning biological age assessment carried out by the Department of Forensic Sciences, Oslo University Hospital, where HA, ØB, KTC and VR are currently employed. PSD, TL, OR, MK and MH declare no potential conflict of interest.

### Additional information

**Supplementary Information** The online version contains supplementary material available at <https://doi.org/10.1038/s41598-023-29381-7>.

**Correspondence** and requests for materials should be addressed to V.R.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2023