

RESEARCH

Open Access



# Refining epigenetic prediction of chronological and biological age

Elena Bernabeu<sup>1</sup> , Daniel L. McCartney<sup>1</sup>, Danni A. Gadd<sup>1</sup>, Robert F. Hillary<sup>1</sup>, Ake T. Lu<sup>2,3</sup>, Lee Murphy<sup>4</sup>, Nicola Wrobel<sup>4</sup>, Archie Campbell<sup>1</sup>, Sarah E. Harris<sup>5</sup>, David Liewald<sup>5</sup>, Caroline Hayward<sup>1,6</sup>, Cathie Sudlow<sup>7,8,9</sup>, Simon R. Cox<sup>5</sup>, Kathryn L. Evans<sup>1</sup>, Steve Horvath<sup>2,3</sup>, Andrew M. McIntosh<sup>1,10</sup>, Matthew R. Robinson<sup>11</sup>, Catalina A. Vallejos<sup>6,12</sup> and Riccardo E. Marioni<sup>1\*</sup>

## Abstract

**Background** Epigenetic clocks can track both chronological age (cAge) and biological age (bAge). The latter is typically defined by physiological biomarkers and risk of adverse health outcomes, including all-cause mortality. As cohort sample sizes increase, estimates of cAge and bAge become more precise. Here, we aim to develop accurate epigenetic predictors of cAge and bAge, whilst improving our understanding of their epigenomic architecture.

**Methods** First, we perform large-scale ( $N = 18,413$ ) epigenome-wide association studies (EWAS) of chronological age and all-cause mortality. Next, to create a cAge predictor, we use methylation data from 24,674 participants from the Generation Scotland study, the Lothian Birth Cohorts (LBC) of 1921 and 1936, and 8 other cohorts with publicly available data. In addition, we train a predictor of time to all-cause mortality as a proxy for bAge using the Generation Scotland cohort (1214 observed deaths). For this purpose, we use epigenetic surrogates (EpiScores) for 109 plasma proteins and the 8 component parts of GrimAge, one of the current best epigenetic predictors of survival. We test this bAge predictor in four external cohorts (LBC1921, LBC1936, the Framingham Heart Study and the Women's Health Initiative study).

**Results** Through the inclusion of linear and non-linear age-CpG associations from the EWAS, feature pre-selection in advance of elastic net regression, and a leave-one-cohort-out (LOCO) cross-validation framework, we obtain cAge prediction with a median absolute error equal to 2.3 years. Our bAge predictor was found to slightly outperform GrimAge in terms of the strength of its association to survival ( $HR_{\text{GrimAge}} = 1.47 [1.40, 1.54]$  with  $p = 1.08 \times 10^{-52}$ , and  $HR_{\text{bAge}} = 1.52 [1.44, 1.59]$  with  $p = 2.20 \times 10^{-60}$ ). Finally, we introduce MethylBrowsR, an online tool to visualise epigenome-wide CpG-age associations.

**Conclusions** The integration of multiple large datasets, EpiScores, non-linear DNAm effects, and new approaches to feature selection has facilitated improvements to the blood-based epigenetic prediction of biological and chronological age.

\*Correspondence:

Riccardo E. Marioni  
riccardo.marioni@ed.ac.uk

Full list of author information is available at the end of the article



© The Author(s) 2023. **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 Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

## Background

The development and application of epigenetic predictors for healthcare research has grown dramatically over the last decade [1]. These predictors can aid disease risk stratification and are based on associations between CpG DNA methylation (DNAm) and age, health, and lifestyle outcomes. DNAm is dynamic, tissue-specific, and is influenced by both genetic and environmental factors. DNAm can precisely track ageing through predictors termed “epigenetic clocks” [2–8]. DNAm has also been found to capture other components of health, such as smoking status [9, 10], alcohol consumption [11, 12], obesity [11, 13], and protein levels [14].

“First generation” epigenetic ageing clocks, including those by Horvath [3] and Hannum et al. [4], were trained on chronological age [2–4] (cAge), with near-perfect clocks expected to arise as sample sizes grow [5]. However, cAge clocks hold limited capability for tracking and quantifying age-related health status, also termed biological age (bAge) [5, 8]. To address this, “second generation” clocks have been trained on other age-related measures, including a phenotypic biomarker of morbidity (PhenoAge [15]), rate of ageing (DunedinPACE [16]), and time to all-cause mortality (GrimAge [17]). Regressing an epigenetic clock predictor (whether trained on cAge or bAge) on chronological age within a cohort gives rise to an “age acceleration” residual with positive values corresponding to faster biological ageing.

Penalised regression approaches such as elastic net [18] are commonly used to derive epigenetic predictors. These identify a weighted linear combination of CpGs that optimally predict an outcome from a statistical perspective, i.e. no preference is given to the location or possible biological role of the input features. The majority consider genome-wide CpG sites as potential predictive features. However, others have used a two-stage approach that first creates DNAm surrogates (or epigenetic scores—EpiScores) for biomarkers (also typically via elastic net) prior to training a second elastic net model on the phenotypic outcome or time to event (TTE) [14, 17]. GrimAge is currently considered one of the best bAge epigenetic clocks [16]. It is derived from age, sex, and EpiScores of smoking pack years and seven plasma proteins that have been associated with mortality or morbidity: adrenomedullin (ADM), beta-2-microglobulin (B2M), cystatin C, growth differentiation factor 15 (GDF15), leptin, plasminogen activation inhibitor 1 (PAI1), and tissue inhibitor metalloproteinase (TIMP1). Recently, a wider set of 109 EpiScores for the circulating proteome were generated by Gadd et al. [14]. These have not yet been considered as potential features for the prediction of bAge.

Here, we sought to improve the prediction of both cAge and bAge (Fig. 1). We first present large-scale

epigenome-wide association studies (EWAS) of cAge (for both linear and quadratic CpG effects) and time to all-cause mortality as a proxy for bAge. A predictor of cAge is then generated using DNAm data from 11 cohorts, including samples from > 18,000 participants of the Generation Scotland study [19]. We use a leave-one-cohort-out (LOCO) prediction framework, including feature pre-selection ahead of elastic net for linear and non-linear DNAm-age relationships (ascertained through the EWAS), to test its performance. Through data linkage to death records in Generation Scotland, we develop a bAge predictor of time to all-cause mortality, which we compare against GrimAge, in four external cohorts. These analyses highlight the potential for large DNAm resources to generate increasingly accurate predictors of (i) cAge, with potential forensic utility, and (ii) bAge, with potential implications for risk prediction and clinical trials.

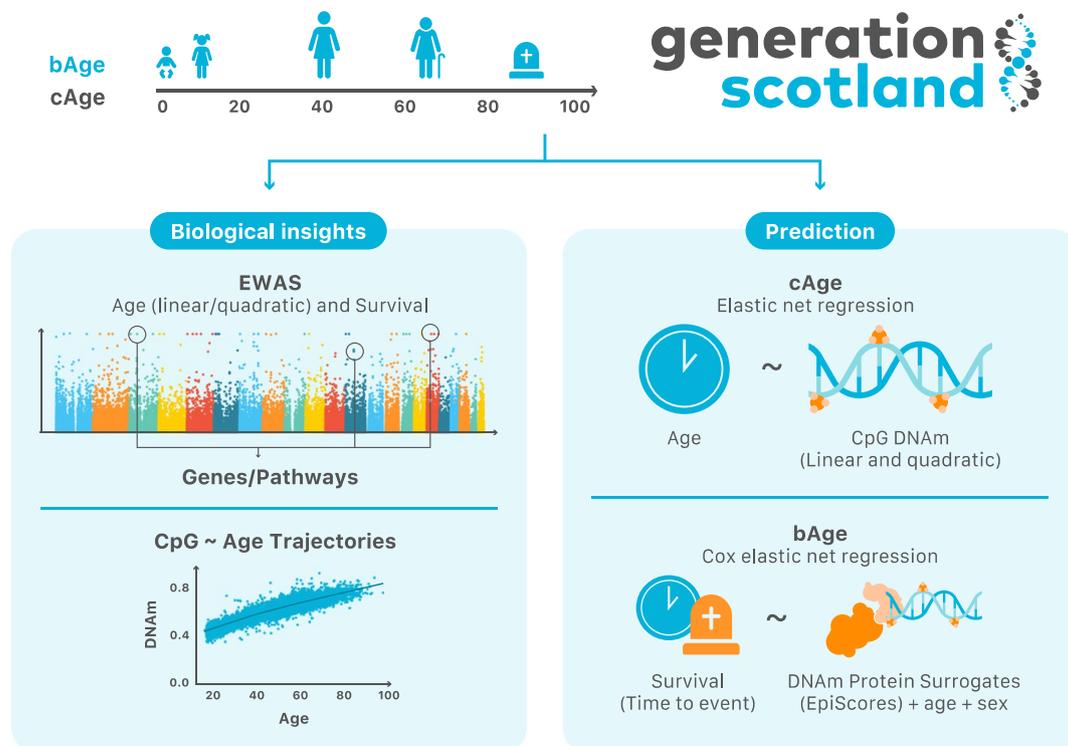
## Methods

### Data overview

Generation Scotland is a Scottish family-based study with over 24,000 participants recruited between 2006 and 2011 [19]. Blood-based DNAm levels at 752,722 CpG sites were quantified using the Illumina MethylationEPIC array for 18,413 individuals. Participants were aged between 18 and 99 years at recruitment, with a mean age of 47.5 years (SD 14.9, Table 1). The data was processed in three sets ( $N_{\text{Set1}} = 5087$ ,  $N_{\text{Set2}} = 4450$ ,  $N_{\text{Set3}} = 8876$ ), with a total of 121 experimental batches (see Additional file 1).

In order to train and test a cAge predictor, Generation Scotland data as well as that from an additional 6261 individuals from ten external cohorts were considered. These included the Lothian Birth Cohorts (LBC) of 1921 and 1936 [20, 21] and eight publicly available Gene Expression Omnibus (GEO) datasets (Table 1) [4, 22–25]. In addition, the independent dataset GEO GSE55763 [13, 26] (2711 samples from 2664 individuals) was used to assess cAge clock performance against existing clocks in individuals not used for training across any of the predictors considered. Given that the external datasets assessed DNAm (blood-based apart from GSE78874, which considered saliva) using the Illumina HumanMethylation450K array, the Generation Scotland data were subset to 374,791 CpGs that were present across all studies. Missing values were mean imputed per CpG and per cohort.

The bAge predictor was trained using data for 18,365 participants from the Generation Scotland cohort for which valid death status data (i.e. death status non-missing, and age at death not lower than age at baseline) via linkage to the National Health Service



**Fig. 1** Study overview. Using the Generation Scotland cohort as our main data source, we explored the relationship between the epigenome and age/survival via EWAS, which also informed on genes of interest and potentially enriched pathways. We further characterised epigenome-wide CpG~age trajectories, which can be visualised in a new Shiny app, MethylBrowsR (<https://shiny.igmm.ed.ac.uk/MethylBrowsR/>). Finally, we refined epigenetic prediction of both cAge and bAge. Calculation of cAge can be performed either using a standalone script ([https://github.com/elena-bernabeu/cage\\_bage/tree/main/cage\\_predictor](https://github.com/elena-bernabeu/cage_bage/tree/main/cage_predictor)) or by uploading DNAm data to our MethylDetectR shiny app (<https://shiny.igmm.ed.ac.uk/MethylDetectR/>). As the weights for GrimAge and its component parts are not publicly available, bAge can only be calculated by using our standalone script ([https://github.com/elena-bernabeu/cage\\_bage/tree/main/bage\\_predictor](https://github.com/elena-bernabeu/cage_bage/tree/main/bage_predictor)), after obtaining GrimAge estimates from an external online calculator (<http://dnamage.genetics.ucla.edu/new>)

Central Register was available. A total of 1214 participant deaths have been recorded as of March 2022, when records were last updated. Alive individuals in March 2022 were censored at their age at that time (TTE thus being age in March 2022 minus age at baseline). Average TTE amongst deaths was 7.79 (SD 3.54) years, and average TTE amongst censored samples was 12.82 (SD 1.35) years. To test the bAge predictor, data from an additional 4134 individuals (with a total of 1653 deaths) from four external cohorts (six datasets) were considered. These included the baseline samples of both the LBC1921 and LBC1936 cohorts, as well as the Framingham Heart Study (FHS) [27–29] and the Women’s Health Initiative (WHI) [30, 31] Broad Agency Award 23 (B23) study for Black, White, and Hispanic individuals (Table 2).

A detailed description of the datasets used (Generation Scotland, GEO, LBC, FHS, and WHI) can be found in Additional file 1.

### Epigenome-wide association study of chronological age

We conducted an EWAS to identify CpG sites that had linear or quadratic associations with chronological age, using Generation Scotland data ( $N=18,413$ , CpGs=752,722). Linear regression analyses were carried out which included both linear and quadratic CpG M-values as independent variables and age as the dependent variable ( $\text{Age} \sim \text{CpG}$  and  $\text{Age} \sim \text{CpG} + \text{CpG}^2$ , respectively). Fixed effect covariates included estimated white blood cell (WBC) proportions (basophils, eosinophils, natural killer cells, monocytes, CD4T, and CD8T cells) calculated in the *minfi* R package (version 1.36.0) [32] using the Houseman method [33], sex, DNAm batch/set, smoking status (a factor with 5 levels: current, gave up in the last year, gave up more than a year ago, never, or unknown), smoking pack years (number of packs of cigarettes smoked per day, 20 cigarettes per pack, multiplied by the number of years the person smoked), and 20 DNAm principal components

**Table 1** Age profile and test set prediction performance for cohorts used in cAge predictor training and testing. Predictions were made using a LOCO approach, where each cohort was excluded in training and the resulting model was used for testing (see [Methods](#)). Models were trained on age, and if an individual was predicted to be under 20, their prediction was re-estimated considering models trained on log(age). External cohort information taken from Zhang et al. [5]. *r* column states Pearson correlation, RMSE the root mean squared error, and MAE the median absolute error

Cohort	N	Mean age (SD)	Age range	N <sub>Females</sub> (%)	Tissue	Prediction accuracy		
						<i>r</i>	RMSE	MAE
GS	18,413	47.5 (14.9)	[17.1, 98.5]	10,833 (58.8%)	Blood	-	-	-
LBC1921 <sup>20,21</sup>	692	82.3 (4.3)	[77.8,90.6]	401 (57.9%)	Blood	0.659	4.050	2.466
LBC1936 <sup>20,21</sup>	2796	73.6 (3.7)	[67.7,80.9]	1356 (48.5%)	Blood	0.685	3.311	2.099
GSE72775 <sup>22</sup>	335	70.2 (10.3)	[36.5, 90.5]	138 (41.2%)	Blood	0.949	3.275	1.843
GSE78874 <sup>22</sup>	259	68.8 (9.7)	[36.0, 88.0]	113 (43.6%)	Saliva	0.875	6.826	4.333
GSE72773 <sup>22</sup>	310	65.6 (13.9)	[35.1, 91.9]	150 (48.4%)	Blood	0.945	4.611	2.068
GSE72777 <sup>22</sup>	46	14.7 (10.4)	[2.2, 35.0]	31 (67.4%)	Blood	0.942	4.211	2.505
GSE41169 <sup>a,23</sup>	95	31.6 (10.3)	[18.0, 65.0]	28 (29.5%)	Blood	0.975	2.869	1.947
GSE40279 <sup>4</sup>	656	64.0 (14.7)	[19.0, 101.0]	338 (51.5%)	Blood	0.969	3.697	2.074
GSE42861 <sup>a,24</sup>	689	51.9 (11.8)	[18.0, 70.0]	492 (71.4%)	Blood	0.972	4.498	3.563
GSE53740 <sup>a,25</sup>	383	67.8 (9.6)	[34.0, 93.0]	155 (40.5%)	Blood	0.921	4.443	2.797

<sup>a</sup> Some cohorts contain case/control data. GSE41169: schizophrenia 62, control 33; GSE42861: rheumatoid arthritis 354, control 335; GSE53740: Alzheimer's disease 15, corticobasal degeneration 1, frontotemporal dementia (FTD) 121, FTD/MND 7, progressive supranuclear palsy 43, control 193, unknown 4

**Table 2** Cox proportional hazards output for GrimAgeAccel and bAgeAccel in the test datasets. Hazard ratios are presented per standard deviation of the GrimAgeAccel and bAgeAccel variables. Further details in Additional File 4: Table S11. Asterisk symbol (\*) indicates the following: the FHS cohort used here was the same as the test set from the original GrimAge paper

Cohort	N	N deaths	GrimAgeAccel Hazard ratio (95% CI)	bAgeAccel Hazard ratio (95% CI)
LBC1936 <sup>20,21</sup>	895	367	1.74 (1.57, 1.94)	1.73 (1.56, 1.91)
LBC1921 <sup>20,21</sup>	421	421	1.33 (1.20, 1.47)	1.44 (1.29, 1.59)
FHS <sup>*27,28</sup>	711	100	1.72 (1.35, 2.19)	1.77 (1.40, 2.25)
WHI B23 White <sup>30,31</sup>	998	418	1.44 (1.31, 1.58)	1.45 (1.32, 1.60)
WHI B23 Black <sup>30,31</sup>	676	229	1.35 (1.19, 1.53)	1.42 (1.24, 1.62)
WHI B23 Hispanic <sup>30,31</sup>	433	118	1.41 (1.18, 1.68)	1.44 (1.21, 1.72)

(PCs) to correct for unmeasured confounders. Family structure was not accounted for given the nature of the phenotype. Age was centred by its mean, and CpG/CpG<sup>2</sup> M-values were scaled to mean zero and variance one. Epigenome-wide significance was set at  $p$ -value  $< 3.6 \times 10^{-8}$ , as per Saffari et al. [34]. For each CpG-age association, *F*-tests were used to compare models including the CpG as a linear term, versus one including both linear and quadratic terms, whilst controlling for all covariates listed here.

### Epigenome-wide association study of time to all-cause mortality

We conducted an EWAS to identify CpG sites (from a total of 752,722 loci) that were associated with time to all-cause mortality in Generation Scotland. Cox proportional hazards (Cox PH) regression models were fit for each CpG site as the predictor of interest using the *coxph* function from the *survival* R package (version 3.3.1), with time to all-cause mortality or censoring as the survival outcome. Fixed effect covariates included those used in the cAge EWAS (age at baseline, sex, batch/set, smoking status, smoking pack years, WBC estimates, and top 20 DNAm PCs). Epigenome-wide significance was set at  $p$ -value  $< 3.6 \times 10^{-8}$ .

To assess whether relatedness in the cohort influenced the results, we fit a Cox PH model with a kinship matrix for each significantly associated CpG, using the *coxme* R package (version 2.2.16).

### Prediction of chronological age

We used elastic net regression to derive a predictor of chronological age from the 374,791 CpG sites common across all cohorts considered in cAge training (description of cohorts in Table 1). The  $L_1$ ,  $L_2$  mixing parameter was set at  $\alpha = 0.5$  based on epigenetic clock precedent [3, 5]. The *biglasso* R package (version 1.5.1) was used [35], with 25-fold cross-validation (CV;  $\sim 1000$  individuals per fold) to select the shrinkage parameter ( $\lambda$ ) that minimised the mean cross-validated prediction error. A sensitivity analysis was performed, assigning individuals from the same

methylation set, batch, and cohort to individual folds, which returned highly similar results.

The effect of including external cohorts in training, as well as accounting for non-linear relationships and pre-selection of features, amongst others, is briefly detailed in Additional file 2. As a result of these analyses, we created a predictor making use of a LOCO framework, training on both log(age) and age, and performing feature pre-selection ahead of elastic net. Here, we describe each of these steps.

**Leave-one-cohort-out**

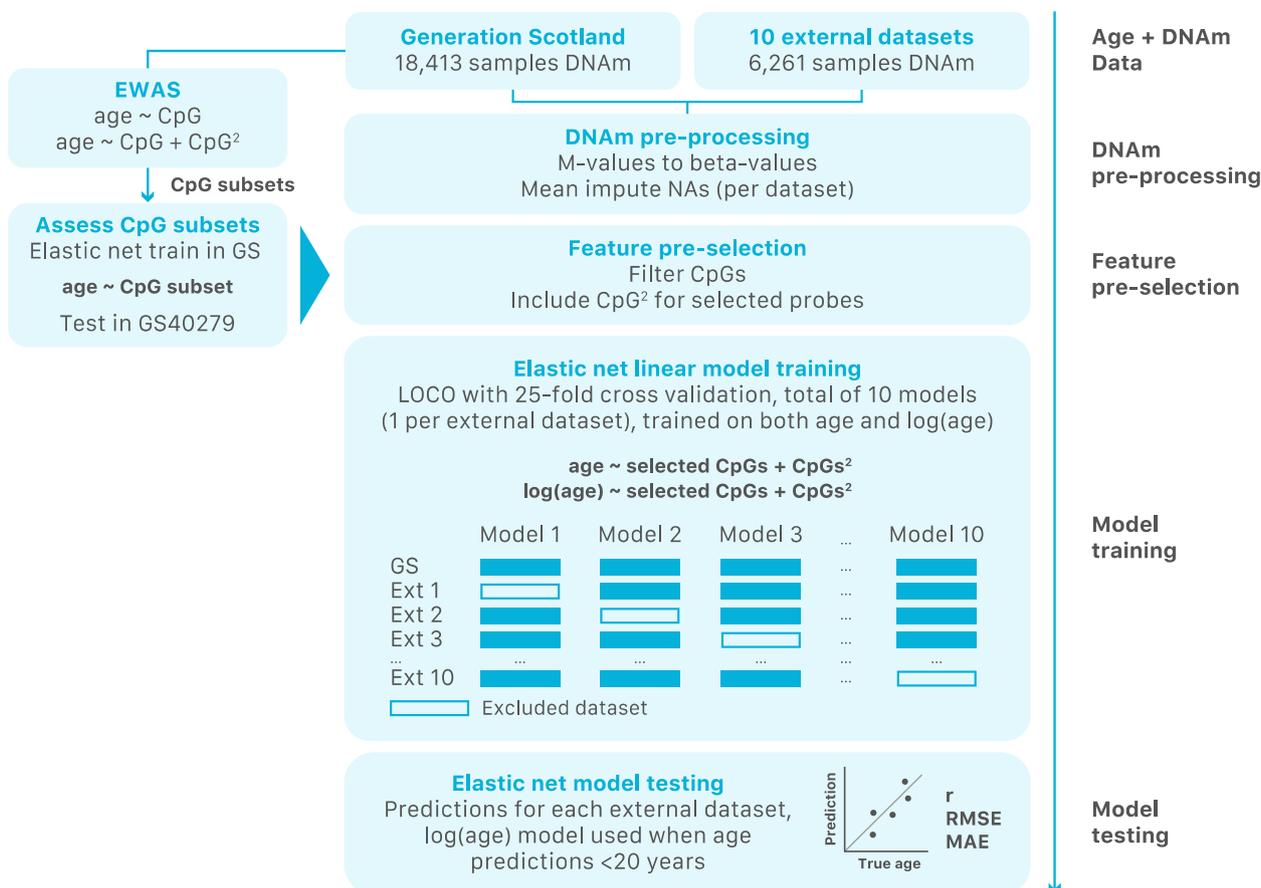
cAge predictors were created using a LOCO framework where, for each of the 10 external cohorts, a model was trained in Generation Scotland and all but one of the external cohorts (Fig. 2). We then tested each of the 10 trained models on the excluded cohort. A final model was trained using all 11 datasets. Pearson correlations (*r*) of cAge predictions with reported age were calculated along with the root mean square error (RMSE) and median absolute error (MAE).

**Log(age)**

In addition to training on chronological age, we also trained models on the natural logarithm of chronological age, log(age), using the same LOCO framework as described above. The age of our test samples was predicted using the model trained on chronological age and, if the value returned was 20 years or younger, a new prediction was obtained making use of the model trained on log(age).

**Feature pre-selection**

Several studies have highlighted the benefits of feature pre-selection for elastic net [36, 37]. Here, we performed preliminary analyses, including differently sized subsets of CpG sites as features in elastic net. After filtering for CpGs present across all datasets (374,791), we considered sites that were epigenome-wide significant at  $p < 3.6 \times 10^{-8}$  and then ranked CpGs in ascending order of *p*-value (most significant ranked first), before defining subsets of varying sizes (from 1000 to 300,000 CpGs). For the purpose of selecting an optimal



**Fig. 2** Flowchart for the creation of the cAge predictor. First, DNAm data originating from Generation Scotland and 10 external datasets was pre-processed. Next, CpGs were pre-selected based on the Generation Scotland EWAS for epigenome-wide significant linear and quadratic CpG-age associations. Elastic net models were then trained and tested on the remaining features using a LOCO framework with 25-fold CV, with training on both age and log(age) as outcomes

number of pre-selected CpGs, we performed a screening using Generation Scotland as our training cohort and GSE40279 (one of the largest external datasets with a wide age range) as our test set. Our analyses showed that the 10,000 most significant loci (age—CpG associations) yielded the test set predictions with the highest  $r$  and lowest RMSE (see [Results](#)). In addition to these sites, subsets of CpGs with a significant quadratic relationship to age were explored, with subset sizes varying from 100 to 20,000. These features were included in training as CpG<sup>2</sup> beta values and, when not already present in the model, in their linear form as well. In addition to the top 10,000 age-associated CpGs, the top 300 quadratic sites from our EWAS yielded the best performing model (see [Results](#)). This final list of features was then used as input for the LOCO framework described above. The final models, trained on all datasets, selected a  $\lambda$  of 0.0308 for the model trained on age and a  $\lambda$  of 0.0006 for the model trained on log(age).

#### **Comparison to ZhangAge, HannumAge, and HorvathAge**

Our final cAge predictor (trained on all 11 datasets in [Table 1](#)) and those by Zhang et al. (ZhangAge) [5], Hannum et al. (HannumAge) [4], and Horvath (HorvathAge) [3] were projected onto the GSE55763 dataset to compare their performance in an independent test set. External clock predictions were calculated using the methylCIPHER R package [38] (<https://github.com/MorganLevineLab/methylCIPHER>).

#### **Prediction of time to all-cause mortality as a proxy for biological age**

##### **Training in Generation Scotland**

To train a bAge predictor, component scores for GrimAge were estimated for all Generation Scotland samples via Horvath's online calculator [17] (<http://dnamage.genetics.ucla.edu/new>). These included EpiScores of smoking and seven proteins—DNAm ADM, DNAm B2M, DNAm cystatin C, DNAm GDF15, DNAm leptin, DNAm PAI1, and DNAm TIMP1. Each variable was then standardised to have a mean of zero and variance of one. We also considered DNAm EpiScores for 109 proteins as described by Gadd et al. [14]. The 109 EpiScores were projected into Generation Scotland via the MethylDetectR [39] Shiny App (<https://shiny.igmm.ed.ac.uk/MethylDetectR/>) before being standardised to have a mean of zero and variance of one.

This resulted in 116 protein EpiScores, a smoking EpiScore, plus chronological age and sex as features for an elastic net Cox PH model (R package *glmnet* version 4.1.4), using time to all-cause mortality or censoring as outcome. A 20-fold CV was performed (with approximately 1000 individuals per fold), with individuals from

the same Generation Scotland technical batch (see [Additional file 1](#)) included in the same fold, and with Harrell's  $C$  index used to identify the optimal  $\lambda$  value (0.0025).

#### **Testing in LBC, FHS, and WHI**

We defined bAge as the weighted linear combination of covariates selected by our Cox PH elastic net model (see [Results](#)). These estimates were then scaled and returned as a predictor with mean of zero and variance of one, for each dataset. A bAgeAccel estimate was also calculated, which is the residual of bAge regressed on chronological age to obtain measure of accelerated epigenetic ageing.

After regressing on age, we assessed the association between our bAge clock, as well as GrimAge, PhenoAge, and DunedinPACE, and time to all-cause mortality in LBC1921 and LBC1936. GrimAge and PhenoAge were calculated using Horvath's online calculator [17], whilst DunedinPACE was calculated via the DunedinPACE R package [16] (<https://github.com/danbelsky/DunedinPACE>). Cox PH models, adjusting for age and sex, were used to evaluate associations between the clocks and all-cause mortality. Further, Cox PH models treating GrimAge, PhenoAge, and DunedinPACE (in turn) as a covariate in addition to our bAge clock were run to assess our predictor's independent association with mortality.

Finally, associations with time to all-cause mortality in four additional external datasets (FHS, and the WHI studies for White, Black, and Hispanic ancestries) were assessed for GrimAge and bAge, the clocks with the largest associations in the LBC cohorts ([Table 2](#)).

We examined Schoenfeld residuals in the LBC1921 and LBC1936 Cox PH models that included age, sex, and our bAge clock as covariates to check the proportional hazards assumption at both global and variable-specific levels using the *cox.zph* function from the R *survival* package (version 3.3.1).

#### **CpG-based predictor of mortality**

We also investigated a direct CpG predictor for time to all-cause mortality (methods and results described in [Additional file 2](#)). This predictor was found to have weaker associations with time to all-cause mortality in the LBC cohorts than the aforementioned bAge estimate, both when training just on CpGs as well as when considering both CpGs and EpiScores as training features.

#### **Enrichment analyses**

Gene set enrichment analyses were performed using the Functional Mapping and Annotation (FUMA) GENE-2FUNC tool [40], which employs a hypergeometric test.

Background genes employed included all unique genes tagged by CpGs in the EPIC array. A false discovery rate (FDR)  $p$ -value threshold was set at 0.05, and the minimum number of overlapping genes within gene sets was set to 2. These analyses did not explicitly account for Illumina chip biases relating to how CpGs are annotated to genes [41], which may have influenced our results.

## Results

### Epigenome-wide association study of chronological age

EWAS of cAge were performed in the Generation Scotland cohort, resulting in 99,832 linear and 137,915 quadratic CpG associations that were epigenome-wide significant ( $p < 3.6 \times 10^{-8}$ , Additional file 3: Figure S1, Additional file 4: Table S1 and S2, see [Methods](#)). These mapped to 17,339 and 19,432 unique genes, respectively. There were 48,312 CpGs with both a significant linear and quadratic association.

The most significant linear associations included cg16867657 and cg24724428 (*ELOVL2*), cg08097417 (*KLF14*), and cg12841266 (*LHFPL4*), all  $p < 1.0 \times 10^{-300}$ , (Additional file 3: Figure S2, Additional file 4: Table S1). Around half of the CpGs with a significant linear association (51,213/99,832, 51.3%) showed a positive association between DNAm and age. The most significant quadratic associations were cg11084334 (*LHFPL4*,  $p = 6.5 \times 10^{-206}$ ), cg15996534 (*LOC134466*,  $p = 8.7 \times 10^{-194}$ ), and cg23527621 (*ECE2* and *CAMK2N2*,  $p = 1.0 \times 10^{-190}$ , Additional file 3: Figure S3, Additional file 4: Table S2).

The univariate associations between all 752,722 CpGs and cAge in a subset of 4450 unrelated participants (DNAm arrays processed together in a single experiment) from Generation Scotland can be visualised via an online ShinyApp, MethylBrowsR (<https://shiny.igmm.ed.ac.uk/MethylBrowsR/>).

### Epigenome-wide association study of time to all-cause mortality

To identify individual CpG loci associated with survival, we performed an EWAS on time to all-cause mortality in Generation Scotland ( $N_{\text{deaths}} = 1214$ ; see [Methods](#)). This analysis identified 1182 epigenome-wide significant associations ( $p < 3.6 \times 10^{-8}$ , Additional file 3: Figure S4), which mapped to 704 unique genes. For around a third (418/1182 = 35.4%) of these CpGs, DNAm was associated with a decreased survival time (HR > 1). The lead findings included CpGs mapping to smoking-related loci [10, 42–46] such as cg05575921 (*AHRR*,  $p = 3 \times 10^{-57}$ ), cg03636183 (*F2RL3*,  $p = 6.8 \times 10^{-44}$ ), cg19859270 (*GPR15*,  $p = 1.1 \times 10^{-33}$ ), cg17739917 (*RARA*,  $p = 1.9 \times 10^{-33}$ ), cg14391737 (*PRSS23*,  $p = 5.6 \times 10^{-33}$ ), cg09935388 (*GFIL*,  $p = 3.3 \times 10^{-31}$ ), and cg25845814 (*ELMSAN1/MIR4505*,

$p = 1.3 \times 10^{-30}$ ) (Additional file 4: Table S3). Amongst the top 50 associations, only one probe has not been previously linked to smoking (assessed via a lookup of findings from the EWAS catalog [47]), cg03546163. This probe maps to *FKBP5*, a gene whose methylation is involved in the regulation of the stress response and which has been linked to increased cardiometabolic risk through accelerated ageing [48]. All associations, except that for cg24364998, remained statistically significant after adjusting for relatedness in the Generation Scotland cohort (see [Methods](#), Additional file 4: Table S4).

There was a high correlation of the Z-score effect sizes across the 200 sites that overlapped between our study and the 257 epigenome-wide significant findings from a recent large ( $N = 12,300$ ,  $N_{\text{deaths}} = 2561$ ) meta-analysis of all-cause mortality [49] ( $r = 0.58$ , Additional file 3: Figure S5). Despite differences in covariate adjustments, all 200 sites were significant at a nominal  $p < 0.05$  threshold, and 25 were epigenome-wide significant at  $p < 3.6 \times 10^{-8}$ .

A gene-set enrichment analysis considering genes to which epigenome-wide significant CpGs mapped returned 198 significantly enriched (FDR  $p < 0.05$ ) GO biological processes (see [Methods](#), full FUMA gene-set enrichment results in Additional file 4: Table S5). The most significantly enriched GO terms included processes relating to neurogenesis/neuron differentiation and development, positive immune system regulation and development, cell motility and organisation, and regulation of protein modification/phosphorylation. Other significantly enriched sets included sites bound by FOXF3, ETS2, and the PML-RARA fusion protein.

### Prediction of chronological age

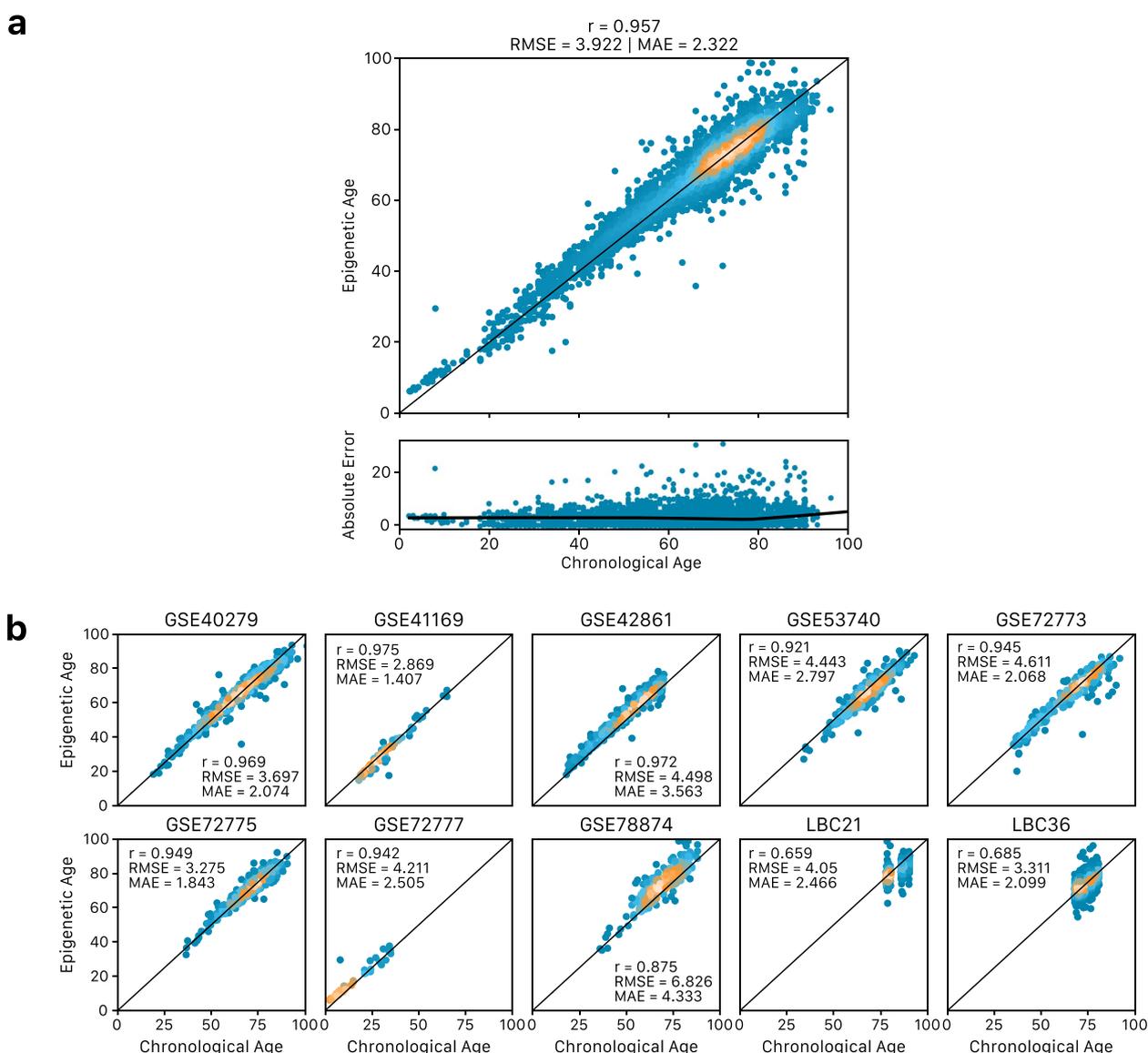
Epigenetic clocks for cAge were created using elastic net penalised regression in a LOCO framework (total of 10 models), with a final cAge clock trained on all data (see [Methods](#), Fig. 2, Additional file 2). In our screening step, after iterating through combinations of CpG and CpG<sup>2</sup> terms (ranked by EWAS  $p$ -value), the best-performing model considered the top 10,000 CpG and top 300 CpG<sup>2</sup> sites from the EWAS as potentially informative features (see [Methods](#), Additional file 3: Figure S6 and S7, Additional file 4: Table S6 and S7). Both age and log(age) were considered as outcomes, with the latter showing better prediction results in younger individuals, reflecting the importance of considering non-linear DNAm-age associations in cAge prediction (see [Methods](#), Additional file 2). As a result, if the initial cAge prediction was < 20 years,

that individual’s predicted age was re-estimated using weights from the log(age) model.

The combined LOCO prediction results (one cAge model per external cohort) showed a strong correlation with cAge ( $r=0.96$ , Fig. 3, Additional file 3: Figure S8, Table 1) and a MAE of 2.3 years. Furthermore, 24% of individuals were classified to within 1 year of their chronological age. The cohort with the largest prediction errors was GSE78874, in which DNAm was measured in saliva instead of blood.

The final cAge predictor (trained in all 11 cohorts) with the lowest mean cross-validated error identified 2330 features (2274 linear and 56 quadratic) as most predictive of age, and 1986 features (1931 linear and 55 quadratic) as most predictive of log(age). The weights for the age model are presented in Additional file 4: Table S8, and for the log(age) model in Additional file 4: Table S9.

Considering a large external cohort ( $N=2711$ ), our cAge predictor ( $r=0.96$ ,  $RMSE=3.04$ ,  $MAE=1.74$ ) outperformed ZhangAge ( $r=0.95$ ,  $RMSE=5.54$ ,  $MAE=3.8$ ),



**Fig. 3** Performance of cAge LOCO framework (one cAge model per external cohort), **a** across all 10 datasets considered, and **b** per cohort. Performance metrics shown include Pearson correlation ( $r$ ), root mean squared error (RMSE), and median absolute error (MAE). Metrics also included in Table 1

HorvathAge ( $r=0.90$ ,  $RMSE=9.1$ ,  $MAE=8.11$ ), and HannumAge ( $r=0.88$ ,  $RMSE=5.08$ ,  $MAE=3.53$ , Fig. 4).

#### Prediction of time to all-cause mortality as a proxy for biological age

In an effort to improve the prediction of bAge, an elastic net Cox model was trained on time to all-cause mortality in Generation Scotland ( $N_{\text{total}}=18,365$ ,  $N_{\text{deaths}}=1214$ ; see [Methods](#)). The GrimAge components (age, sex, and EpiScores for smoking and 7 plasma proteins) and Gadd et al.'s 109 protein EpiScores [14] were considered as potentially-informative features (Fig. 5).

The elastic net Cox PH model identified a weighted sum of 35 features as most predictive of time to all-cause mortality in Generation Scotland. These included age and the GrimAge smoking EpiScore, along with 5/7 protein EpiScores from GrimAge (B2M, cystatin C, GDF15, PAI1, and TIMP1), and 28/109 protein EpiScores from Gadd et al. [14]. Amongst these were EpiScores for C-reactive protein (CRP), the growth hormone receptor (GHR) protein, and numerous cytokines (CCL11, CCL23, CCL18, CXCL10, CXCL9, CXCL11, and HGF). The weights for the linear predictor are presented in Additional file 4: Table S10.

Our bAge predictor was regressed on age to obtain a measure of epigenetic age acceleration (bAgeAccel). The epigenetic age acceleration residuals showed significant associations with time to all-cause mortality across six test datasets of differing ancestries (Table 2, Additional file 4: Table S11, Fig. 6).

We assessed our predictor's association with time to all-cause mortality in comparison to three other epigenetic clocks: GrimAge, PhenoAge, and DunedinPACE (age acceleration residuals after regressing the clock estimates on age) in the LBC1921 and LBC1936 cohorts (see [Methods](#)). Our bAge predictor showed stronger associations (in magnitude and statistical significance) with time to all-cause mortality than

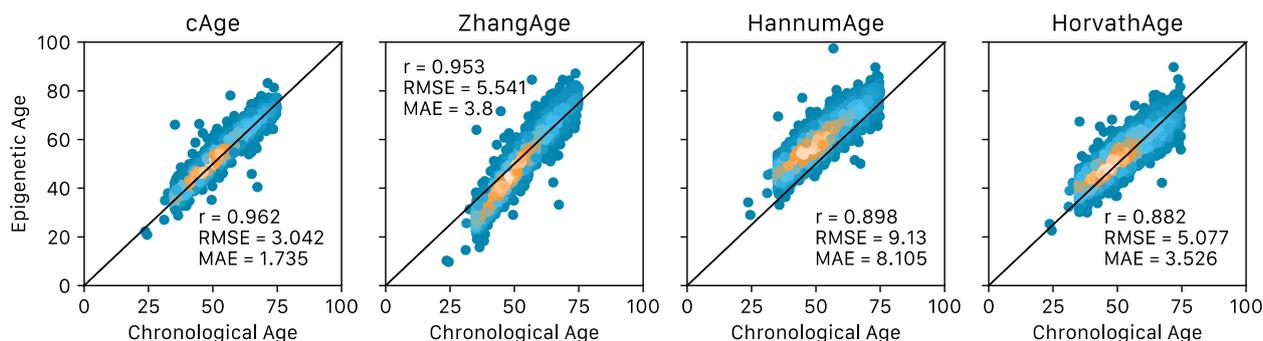
DunedinPACE and PhenoAge and similar performance to GrimAge (Additional file 3: Figure S9, Additional file 4: Table S12). Our bAge predictor's association with time to all-cause mortality remained significant ( $p<0.05$ ) after adjusting for GrimAge, PhenoAge, and DunedinPACE as covariates in three separate models (Additional file 4: Table S13). Considering all six test datasets, the bAge measure showed slightly stronger associations than GrimAge in fixed effects meta-analyses (HR and 95% confidence interval per SD difference of GrimAgeAccel and bAgeAccel: HR = 1.47 [1.40, 1.54] with  $p=1.08 \times 10^{-52}$ , and HR = 1.52 [1.44, 1.59] with  $p=2.20 \times 10^{-60}$ , respectively (Table 2, Additional file 4: Table S11, Fig. 6).

Schoenfeld residual analyses highlighted violations to the proportional hazards assumption at global and variable specific levels for LBC1921 and LBC1936. However, re-running the analysis with different TTE censoring (thresholding at each possible integer year of follow-up) showed minimal differences in the bAgeAccel-survival HRs between models that did not violate the assumption and those that did (see [Methods](#), Additional file 4: Table S14).

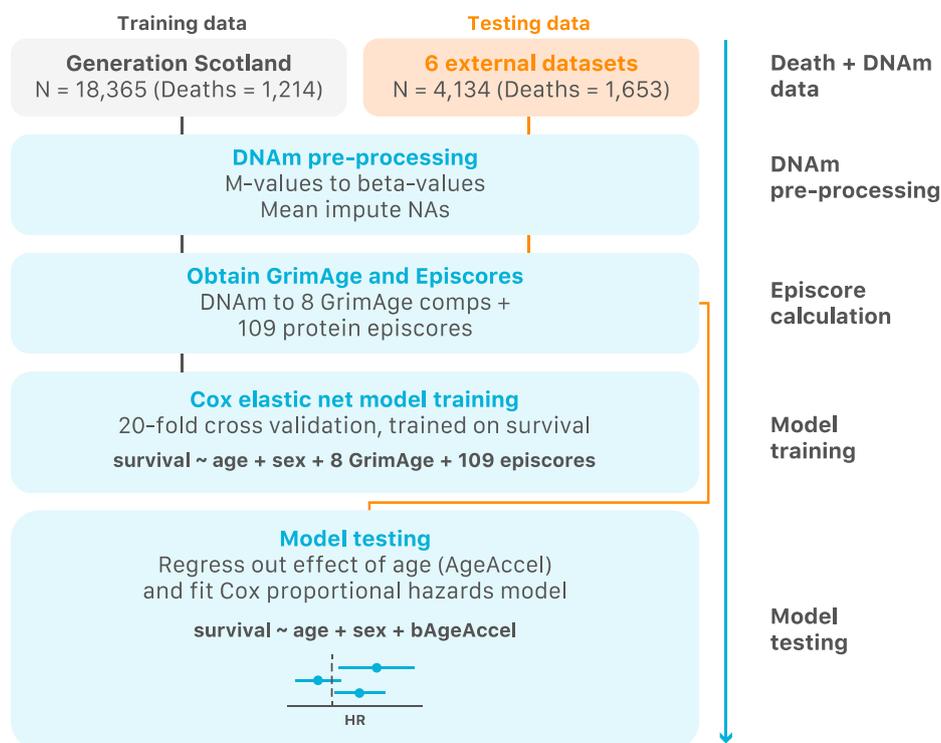
#### Discussion

Accurate predictors of cAge and bAge have major implications for biomedical science and healthcare through risk prediction and preventative medicine. Here, we present improved DNAm-based predictors of age and lifespan.

Epigenetic cAge prediction is expected to reach near-perfect estimates as sample sizes grow [5]. Making use of Generation Scotland, a very large single-cohort DNAm resource, we derived a cAge predictor with a MAE of 1.7 years, tested in over 2000 external samples (Fig. 4). Our predictor has potential forensic applications, although ethical caveats exist [8]. In addition, despite the high correlations and low RMSE and MAE estimates



**Fig. 4** cAge predictor performance in the GSE55763 dataset, compared to ZhangAge, HannumAge, and HorvathAge. Performance metrics shown include Pearson correlation ( $r$ ), root mean squared error (RMSE), and median absolute error (MAE)



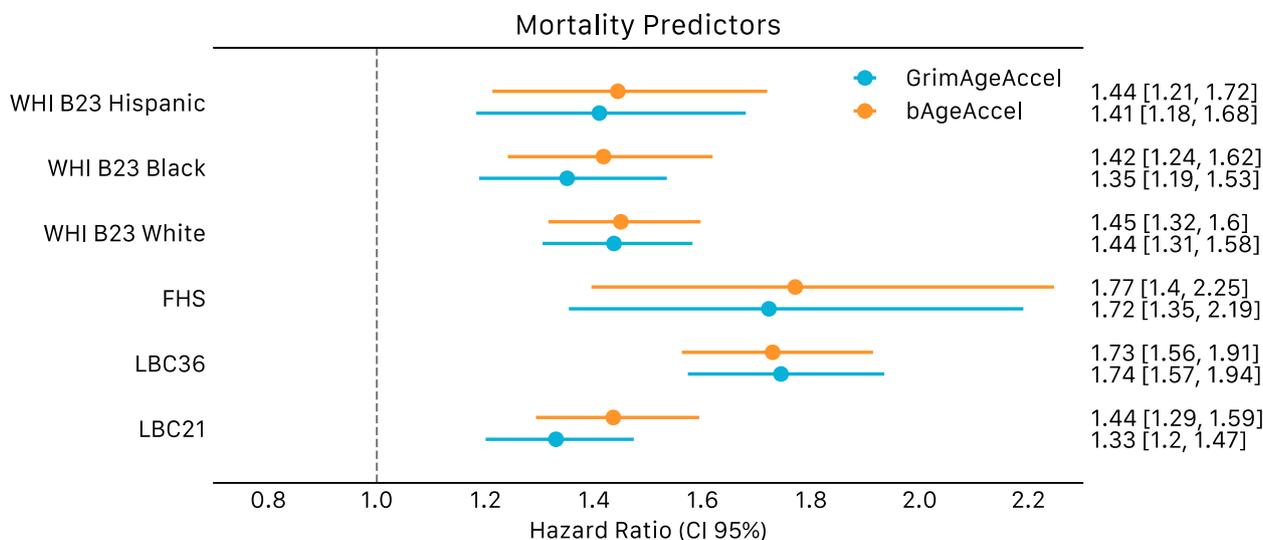
**Fig. 5** Flowchart for the creation of the bAge predictor. First, DNAm data originating from Generation Scotland and six external datasets was pre-processed. GrimAge components and 109 protein EpiScores were generated within each cohort. A Cox PH elastic net regression model of time to all-cause mortality (with 20-fold CV) was trained in Generation Scotland with the GrimAge components and EpiScores as possible features. The model that maximised Harrell’s C index was tested on the six external datasets

at the population level, there are still several individuals with inaccurate predictions (e.g. >20 years between predicted and actual age, Fig. 3), though this could also reflect sample mix-ups or data entry errors.

cAge prediction was improved when accounting for non-linear relationships between DNAm and age (Additional file 2, Additional file 3: Figure S7). Whilst generally understudied, non-linear patterns have been found at numerous CpG sites, where DNAm is found to increase rapidly in early ages and stabilise in adulthood, potentially reflecting developmental processes [50]. Similarly, stable DNAm levels followed by rapid methylation/demethylation have also been described in later life [51], which could offer insight into ageing-specific processes. Given the number of samples from individuals aged 20 or under in the training of our predictor ( $N = 574/24,674 = 2.4\%$ ), we may not have captured the full extent of DNAm-based ageing patterns in the younger population. Future studies could also consider sex-specific models, as diverging non-linear patterns between males and females have been shown previously [52]. Interactions between CpGs along with higher order polynomial terms and spline-based models might better capture some of these non-linear changes.

The development of the cAge predictor highlighted the advantages of feature pre-selection ahead of penalised elastic net regression. Compared to a model with all possible features in the training set ( $r = 0.93$ ,  $RMSE = 5.25$ ,  $MAE = 3.43$ , see Additional file 2), pre-selection greatly improved performance ( $r = 0.96$ ,  $RMSE = 3.92$ ,  $MAE = 2.32$ , Fig. 3). Several DNAm studies of age and age-related phenotypes have used pre-selection methods (e.g. filtering by magnitude of correlation or strength of association) instead of, or in addition to elastic net [53–60]. Whereas the feature pre-selection here required arbitrary decisions on thresholds, other studies have found that feature reduction via PCA optimises DNAm predictors [36, 37].

Feature pre-selection may have aided cAge predictions by screening out CpGs with low intra-sample variability due to technical variance [61, 62]. One previous study [37] observed that CpGs with stronger cAge associations were more reliable. A limitation of our approach to feature pre-selection was that it was biased towards the Generation Scotland cohort in which the age EWAS were conducted. We also note that pre-selection introduces statistical challenges associated with post-selection inference [63]. Furthermore, our penalised regression modelling strategy for



**Fig. 6** Forest plots of bAge/GrimAge predictors, applied to time to all-cause mortality in LBC1921, LBC1936, FHS, and WHI. Predictors regressed on age. Hazard ratios are presented per standard deviation of the GrimAgeAccel and bAgeAccel variables, along with 95% confidence intervals. Cox models are adjusted for age at DNAm sampling and sex

cAge only incorporated additive effects. Non-additive tree ensemble methods and other machine learning frameworks may improve predictions further [64]. Finally, as our predictor has been mainly trained and tested on blood data, it may not generalise to other tissues.

Whilst a single DNAm predictor of cAge is of interest, the selected CpG features are unlikely to identify all epigenome-wide patterns related to ageing. Our EWAS of chronological age identified 99,832 linear and 137,915 quadratic CpG-age associations. The sample size was more than double that of the largest study reported on the EWAS Catalog [47]—our previous Generation Scotland analysis [65]. In addition to refining our previously described DNAm-age linear associations, we have extended previous small-scale approaches to highlight non-linear patterns [51, 52]. As shown here, these findings can aid the predictive performance of epigenetic clocks and may additionally improve our understanding of epigenetic changes during development and ageing-related decline in later life.

Recent work has shifted focus from the prediction of cAge to bAge, with more expansive clinical applications. Our new bAge predictor of all-cause mortality had a greater effect size and was more statistically significant than GrimAge in the external test set meta-analysis. GrimAge is already being used as an end-point for clinical trials [66] and studies of rejuvenation [67, 68]. Our bAge predictor includes five of the seven original GrimAge EpiScores, with ADM and leptin not being selected as features. In addition, it includes 28 protein EpiScores

from Gadd et al. [14]. Amongst the additional protein EpiScores selected by our predictor were those for CRP and numerous cytokines, which reflect inflammation and predict overall and cardiovascular mortality [69–71]. Chronic inflammation can lead to several diseases, including cardiovascular disease, and exacerbates the ageing process [72, 73]. In addition, the growth hormone receptor (GHR) protein EpiScore was selected; both the receptor and its corresponding protein have been linked to longevity in mouse models [74–78]. Twenty-five of the 28 of the selected EpiScores from Gadd et al. [14] have been associated to multiple diseases, including diabetes, chronic obstructive pulmonary disease, ischaemic heart disease, lung cancer, Alzheimer’s, rheumatoid arthritis, stroke, and depression (Additional file 4: Table S10). As sample sizes for cause-specific mortality outcomes increase, a more granular suite of lifespan predictors can be developed. Future studies may also consider the cost implications of profiling thousands of CpGs against the potential improvements in health-span and savings from delaying or preventing disease.

Whereas the cAge predictor is directly applicable and interpretable for a new individual, bAge estimates are relative to the values of other participants in the testing dataset, given the within-cohort scaling of the input features prior to projection. Reporting findings per SD of bAgeAccel will therefore help to facilitate cross-cohort comparisons. Future work for these (and all) DNAm array-based predictors should consider the limitations of signatures that lack absolute thresholds/cut-points for

risk prediction in a new individual selected at random from the population.

A total of 1182 epigenome-wide significant associations were identified in our EWAS of all-cause mortality. The most significant probes mapped to genes previously associated with smoking, such as *AHRR*, *F2RL3*, and *GPR15* [79]. Hypomethylation at probes nearby these genes has been previously linked to increased mortality risk, be that all-cause or disease specific (e.g. cancer or cardiovascular-related mortality) [29, 42, 80, 81]. There was moderate agreement (correlation of 0.58 between Z-scores) between our findings and the significant results from a previous EWAS meta-analysis of survival. However, different covariates and ancestries were considered across these studies. An enrichment analysis highlighted links to neurodevelopment and immune regulation, as well as to sites bound by *FOXP3*, *ETS2*, and the *PML-RARA* fusion protein. *FOXP3* is a transcriptional regulator involved in the development and inhibitory function of regulatory T cells [82]. *ETS2* and *PML-RARA* are a protooncogene and a protein resulting from a chromosomal translocation that generates an oncofusion protein, respectively, having both been linked to acute myeloid leukemia [83, 84]. This finding may be influenced by the large number of cancer-related deaths in Generation Scotland ( $N=509$ ). Further work is needed to disentangle the role of methylation/demethylation at these sites with survival, including the fitting of models with more complete sets of comorbidities, risk, and lifestyle factors. Future EWAS on specific mortality causes will highlight mechanisms underlying age- and disease-related decline.

Importantly, the majority of Generation Scotland participants are of White British ancestry, meaning analyses could present biases towards this population. Whilst our cAge predictor, which was trained on Generation Scotland and external cohorts of multiple ancestries (White, Hispanic, South Asian, East Asian), showed similar accuracy across all testing datasets (Fig. 3b), the magnitude of the survival effect size for bAge was slightly reduced (though still statistically significant) when considering African American and Hispanic ancestry samples, as opposed to European American samples in the WHI cohort (Fig. 6, Table 2). Additionally, our EWASs of survival and age were conducted only using Generation Scotland data. In this context, large multi-ancestry and multi-omic cohorts are needed.

## Conclusions

The integration of multiple large datasets and new approaches to feature selection has facilitated improvements to the blood-based epigenetic prediction of biological and chronological age. The inclusion of multiple

protein EpiScore features and consideration of quadratic DNAm effects may also be relevant for other EWAS and prediction studies. Together, this can improve our biological understanding of complex traits and the prediction of adverse health outcomes.

## Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s13073-023-01161-y>.

**Additional file 1.** Dataset descriptions for Generation Scotland, LBC1921 and LBC1936, GEO datasets, FHS, and WHI. (PDF 88Kb).

**Additional file 2.** Additional cAge and bAge predictor analyses that informed the decisions made in the creation of the latter as described in this manuscript. (PDF 3.3Mb).

**Additional file 3. Figure S1.** Manhattan plots of linear and quadratic age EWAS in Generation Scotland. **Figure S2.** Scatterplots of top 15 associations from age ~ CpG EWAS. **Figure S3.** Scatterplots of top 15 associations from the age ~ CpG + CpG<sup>2</sup> EWAS. **Figure S4.** Manhattan plot of all-cause mortality EWAS. **Figure S5.** Comparison of Z-values for 200 overlapping epigenome-wide significant CpG-mortality associations reported by Colicino et al., and those considered in the present study. **Figure S6.** Prediction metrics for GSE40279, as a function of CpGs included in training of a cAge predictor, trained using Generation Scotland data. **Figure S7.** Prediction metrics for GSE40279, as a function of CpG<sup>2</sup>s included in training of a cAge predictor trained using Generation Scotland data, in addition to top 10K age-associated CpGs. **Figure S8.** cAge predictor median absolute error (MAE) +/- 1.96 its standard deviation on 10 external testing datasets, as a function of age, in 10 year intervals. **Figure S9.** Forest plots of bAge, GrimAge, PhenoAge, and DunedinPACE predictors, applied to all-cause mortality in LBC1921 and LBC1936. (PDF 12.2Mb).

**Additional file 4. Table S1.** Top 10,000 epigenome-wide significant associations between CpG and age. **Table S2.** Top 10,000 epigenome-wide significant associations between CpG<sup>2</sup> and age. **Table S3.** Epigenome-wide significant CpG associations with all-cause mortality. **Table S4.** Epigenome-wide significant CpG associations with all-cause mortality, replicated using the *coxme* R package to account for relatedness. **Table S5.** Gene-set enrichment results for mortality EWAS hits, as returned by FUMA. **Table S6.** Model performance as a function of CpG subset size input into elastic net. **Table S7.** Model performance as a function of CpG<sup>2</sup> subset size input into elastic net, in addition to top 10K linear CpG-age associated sites. **Table S8.** Components of cAge predictor, along with their respective weights. **Table S9.** Components of log(cAge) predictor, along with their respective weights. **Table S10.** bAge predictor components, along with their weights. **Table S11.** bAgeAccel and GrimAgeAccel associations to mortality across 6 datasets. **Table S12.** bAgeAccel, GrimAgeAccel, PhenoAgeAccel, and DunedinPACEAccel associations to mortality across 2 datasets. **Table S13.** bAgeAccel associations to mortality across 2 datasets, including another biological age clock as a covariate (GrimAgeAccel, PhenoAgeAccel, or DunedinPACEAccel). **Table S14.** Schoenfeld residuals and HR as a function of TTE cut-off in bAge prediction. (XLSX 5.3Mb).

## Acknowledgements

We are grateful to all the families who took part, the general practitioners, and the Scottish School of Primary Care for their help in recruiting them and the whole GS team that includes interviewers, computer and laboratory technicians, clerical workers, research scientists, volunteers, managers, receptionists, healthcare assistants, and nurses.

## Authors' contributions

E.B., D.L.M., and A.T.L. analysed the data. D.L.M. developed MethylBrowsR. D.A.G., R.F.H., L.M., N.W., A.C., S.E.H., D.L., C.H., C.S., S.R.C., K.L.E., and A.M.M. were involved in the data generation. E.B., R.E.M. and C.A.V. drafted the initial manuscript. E.B., M.R.R., C.A.V., and R.E.M. designed the study. All authors read and approved the final manuscript.

## Funding

**Generation Scotland:** Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorate (CZD/16/6) and the Scottish Funding Council (HR03006). Genotyping and DNA methylation profiling of the Generation Scotland samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, Edinburgh, Scotland, and was funded by the Medical Research Council UK and the Wellcome Trust (Wellcome Trust Strategic Award Stratifying Resilience and Depression Longitudinally (STRADL; Reference 104036/Z/14/Z). The DNA methylation data assayed for Generation Scotland was partially funded by a 2018 NARSAD Young Investigator Grant from the Brain & Behavior Research Foundation (Ref: 27404; awardee: Dr David M Howard) and by a JMAS SIM fellowship from the Royal College of Physicians of Edinburgh (Awardee: Dr Heather C Whalley).

**Lothian Birth Cohorts:** We thank the LBC1921 and LBC1936 participants and team members who contributed to these studies. The LBC1921 was supported by the UK's Biotechnology and Biological Sciences Research Council (BBSRC), The Royal Society, and The Chief Scientist Office of the Scottish Government. The LBC1936 is supported by the BBSRC, and the Economic and Social Research Council [BB/W008793/1] (which supports S.E.H.), Age UK (Disconnected Mind project), the Medical Research Council (MR/M01311/1), and the University of Edinburgh. Methylation typing of LBC1936 was supported by the Centre for Cognitive Ageing and Cognitive Epidemiology (Pilot Fund award), Age UK, The Wellcome Trust Institutional Strategic Support Fund, The University of Edinburgh, and The University of Queensland. Genotyping was funded by the BBSRC (BB/F019394/1). S.R.C. is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 221890/Z/20/Z).

D.A.G. is supported by funding from the Wellcome Trust 4 year PhD in Translational Neuroscience: training the next generation of basic neuroscientists to embrace clinical research [I08890/Z/15/Z]. R.F.H. is supported by an MRC IEU Fellowship. M.R.R. was funded by Swiss National Science Foundation Eccellenza Grant PCEGP3-181,181 and by core funding from the Institute of Science and Technology Austria. C.H. is supported by an MRC Human Genetics Unit programme grant 'Quantitative traits in health and disease' (U. MC\_UU\_00007/10). E.B. and R.E.M. are supported by Alzheimer's Society major project grant AS-PG-19b-010.

This research was funded in whole, or in part, by the Wellcome Trust (104,036/Z/14/Z, 108,890/Z/15/Z, and 221,890/Z/20/Z). For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

## Availability of data and materials

According to the terms of consent for Generation Scotland participants, access to data must be reviewed by the Generation Scotland Access Committee. Applications should be made to [access@generationscotland.org](mailto:access@generationscotland.org).

Lothian Birth Cohort data are available on request from the Lothian Birth Cohort Study, University of Edinburgh (<https://www.ed.ac.uk/lothian-birth-cohorts/data-access-collaboration>). Lothian Birth Cohort data are not publicly available due to them containing information that could compromise participant consent and confidentiality.

All custom R (version 4.0.3), Python (version 3.9.7), and bash code is available with open access at the following GitHub repository [85]: [https://github.com/elenabernabeu/cage\\_bage](https://github.com/elenabernabeu/cage_bage)

EWAS summary statistics are available on Edinburgh DataShare, along with mean and SD of CpG *M*-values and *F*-test results [86]: <https://datashare.ed.ac.uk/handle/10283/4781> (<https://doi.org/10.7488/ds/3792>).

cAge predictions can be obtained using MethylDetectR (<https://shiny.igmm.ed.ac.uk/MethylDetectR/>) or via a standalone script:

[https://github.com/elenabernabeu/cage\\_bage/tree/main/cage\\_predictor](https://github.com/elenabernabeu/cage_bage/tree/main/cage_predictor)

As the CpG weights for the GrimAge components are not publicly available, bAge predictions first require users to generate GrimAge estimates from the following online calculator (<http://dnamage.genetics.ucla.edu/new>). bAge can then be estimated via the following standalone script: [https://github.com/elenabernabeu/cage\\_bage/tree/main/bage\\_predictor](https://github.com/elenabernabeu/cage_bage/tree/main/bage_predictor)

Visualization of CpG-age relationships can be viewed using MethylBrowsR: <https://shiny.igmm.ed.ac.uk/MethylBrowsR/>

## Declarations

### Ethics approval and consent to participate

All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). All participants provided broad and enduring written informed consent for biomedical research. Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 15/0040/ES), providing generic ethical approval for a wide range of uses within medical research. This study was performed in accordance with the Helsinki declaration.

Ethical approval for the LBC1921 and LBC1936 studies was obtained from the Multi-Centre Research Ethics Committee for Scotland (MREC/01/0/56) and the Lothian Research Ethics committee (LREC/1998/4/183; LREC/2003/2/29). In both studies, all participants provided written informed consent. These studies were performed in accordance with the Helsinki declaration.

### Consent for publication

Not applicable.

### Competing interests

R.E.M. has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock Development Foundation. R.F.H. has received consultant fees from Illumina. R.E.M., R.F.H., and D.A.G. have received consultant fees from Optima partners. A.M.M. has previously received speaker fees from Janssen and Illumina and research funding from The Sackler Trust. M.R.R. receives research funding from Boehringer Ingelheim. All other authors declare no competing interests.

### Author details

<sup>1</sup>Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. <sup>2</sup>Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA. <sup>3</sup>Altos Labs, San Diego, USA. <sup>4</sup>Edinburgh Clinical Research Facility, University of Edinburgh, Edinburgh, UK. <sup>5</sup>Department of Psychology, Lothian Birth Cohorts, University of Edinburgh, Edinburgh, UK. <sup>6</sup>Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. <sup>7</sup>Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK. <sup>8</sup>BHF Data Science Centre, Health Data Research UK, London, UK. <sup>9</sup>Edinburgh Medical School, Usher Institute, University of Edinburgh, Edinburgh, UK. <sup>10</sup>Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK. <sup>11</sup>Institute of Science and Technology Austria, Klosterneuburg, Austria. <sup>12</sup>The Alan Turing Institute, London, UK.

Received: 8 September 2022 Accepted: 6 February 2023

Published online: 28 February 2023

## References

1. Yousefi PD, et al. DNA methylation-based predictors of health: applications and statistical considerations. *Nat Rev Genet.* 2022;23:369–83.
2. Bocklandt S, et al. Epigenetic predictor of age. *PLoS One.* 2011;6:e14821.
3. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol.* 2013;14:1–20.
4. Hannum G, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell.* 2013;49:359–67.
5. Zhang Q, et al. Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* 2019;11:1–11.
6. Field AE, et al. DNA methylation clocks in aging: categories, causes, and consequences. *Mol Cell.* 2018;71:882.
7. Horvath S, Raj K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat Rev Genet.* 2018;19(6):371–84.
8. Bell CG, et al. DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* 2019;20(1):249.
9. McCartney DL, et al. Epigenetic signatures of starting and stopping smoking. *EBioMedicine.* 2018;37:214–20.
10. Joehanes R, et al. Epigenetic signatures of cigarette smoking. *Circ Cardiovasc Genet.* 2016;9:436–47.

11. McCartney DL, et al. Epigenetic prediction of complex traits and death. *Genome Biol.* 2018;19:136.
12. Liu C, et al. A DNA methylation biomarker of alcohol consumption. *Mol Psychiatry.* 2018;23:422–33.
13. Wahl S, et al. Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature.* 2017;541:81–6.
14. Gadd DA, et al. Epigenetic scores for the circulating proteome as tools for disease prediction. *Elife.* 2022;11:e71802.
15. Levine ME, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany NY).* 2018;10:573–91.
16. Belsky DW, et al. DunedinPACE, A DNA methylation biomarker of the pace of aging. *Elife.* 2022;11:e73420.
17. Lu AT, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany NY).* 2019;11:303–27.
18. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B (Statistical Methodol).* 2005;67:301–20.
19. Smith BH, et al. Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS) the study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol.* 2013;42:689–700.
20. Deary IJ, Gow AJ, Pattie A, Starr JM. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *Int J Epidemiol.* 2012;41:1576–84.
21. Taylor AM, Pattie A, Deary IJ. Cohort Profile Update: the Lothian Birth Cohorts of 1921 and 1936. *Int J Epidemiol.* 2018;47:1042–60.
22. Horvath S, et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol.* 2016;17:1–23.
23. Horvath S, et al. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* 2012;13:R97.
24. Liu Y, et al. Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis. *Nat Biotechnol.* 2013;31:142–7.
25. Li Y, et al. An epigenetic signature in peripheral blood associated with the haplotype on 17q21.31, a risk factor for neurodegenerative tauopathy. *PLOS Genet.* 2014;10:e1004211.
26. Lehne B, et al. A coherent approach for analysis of the Illumina HumanMethylation450 BeadChip improves data quality and performance in epigenome-wide association studies. *Genome Biol.* 2015;16:37.
27. Dawber TR, Meadors GF, Moore FE. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health.* 1951;41:279–81.
28. Kannel WB, Feinleib M, Mcnamara PM, Garrison RJ, Castelli WP. An investigation of coronary heart disease in families. the Framingham offspring study. *Am J Epidemiol.* 1979;110:281–90.
29. Philibert RA, Dogan MV, Mills JA, Long JD. AHRH methylation is a significant predictor of mortality risk in Framingham Heart Study. *J Ineur Med.* 2019;48:79–89.
30. Anderson G, et al. Design of the Women's Health Initiative clinical trial and observational study. *Control Clin Trials.* 1998;19:61–109.
31. Anderson GL, et al. Implementation of the women's health initiative study design. *Ann Epidemiol.* 2003;13:55–17.
32. Aryee MJ, et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics.* 2014;30:1363.
33. Houseman EA, et al. DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics.* 2012;13:1–16.
34. Saffari A, et al. Estimation of a significance threshold for epigenome-wide association studies. *Genet Epidemiol.* 2018;42:20–33.
35. Zeng Y, Breheny P. The biglasso package: a memory- and computation-efficient solver for lasso model fitting with big data in R. *R J.* 2017;12:1–14.
36. Doherty, T. et al. A comparison of feature selection methodologies and learning algorithms in the development of a DNA methylation-based telomere length estimator. *bioRxiv.* 2022.04.02.486242 (2022). <https://doi.org/10.1101/2022.04.02.486242>
37. Higgins-Chen AT, et al. A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking. *Nat Aging.* 2022;2022:1–18. <https://doi.org/10.1038/s43587-022-00248-2>.
38. Thrush, K. L., Higgins-Chen, A. T., Liu, Z. & Levine, M. E. R methylCIPHER: a methylation clock investigational package for hypothesis-driven evaluation & research. *bioRxiv.* 2022.07.13.499978 (2022). <https://doi.org/10.1101/2022.07.13.499978>
39. Hillary RF, Marioni RE. MethylDetectR: a software for methylation-based health profiling. *Wellcome Open Res.* 2021;5:283.
40. Watanabe K, Taskesen E, Van Bochoven A, Posthuma D. Functional mapping and annotation of genetic associations with FUMA. *Nat Commun.* 2017;8(1):1826.
41. Maksimovic J, Oshlack A, Phipson B. Gene set enrichment analysis for genome-wide DNA methylation data. *Genome Biol.* 2021;22:173.
42. Bojesen SE, Timpson N, Relton C, Davey Smith G, Nordestgaard BGAHRR. (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax.* 2017;72:646–53.
43. Zhang Y, Yang R, Burwinkel B, Breitling LP, Brenner H. F2RL3 methylation as a biomarker of current and lifetime smoking exposures. *Environ Health Perspect.* 2014;122:131–7.
44. Haase T, et al. Novel DNA methylation sites influence GPR15 expression in relation to smoking. *Biomolecules.* 2018;8:74.
45. Guida F, et al. Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum Mol Genet.* 2015;24:2349–59.
46. Sun YQ, et al. Assessing the role of genome-wide DNA methylation between smoking and risk of lung cancer using repeated measurements: the HUNT study. *Int J Epidemiol.* 2021;50:1482.
47. Battram T, et al. The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res.* 2022;7:41.
48. Beach, S. R. H. et al. Methylation of FKBP5 is associated with accelerated DNA methylation ageing and cardiometabolic risk: replication in young-adult and middle-aged Black Americans. (2021). <https://doi.org/10.1080/15592294.2021.1980688>
49. Colicino E, et al. Blood DNA methylation sites predict death risk in a longitudinal study of 12,300 individuals. *Aging (Albany NY).* 2020;12:14092–124.
50. Alisch RS, et al. Age-associated DNA methylation in pediatric populations. *Genome Res.* 2012;22:623–32.
51. Johnson ND, et al. Non-linear patterns in age-related DNA methylation may reflect CD4+ T cell differentiation. *Epigenetics.* 2017;12:492.
52. Vershina O, Bacalini MG, Zaikin A, Franceschi C, Ivanchenko M. Disentangling age-dependent DNA methylation: deterministic, stochastic, and nonlinear. *Sci Reports.* 2021;11(1):9201.
53. Koch CM, Wagner W. Epigenetic-aging-signature to determine age in different tissues. *Aging (Albany NY).* 2011;3:1018–27.
54. Karir P, Goel N, Garg VK. Human age prediction using DNA methylation and regression methods. *Int J Inf Technol.* 2020;12:373–81.
55. Bekaert B, Kamalandua A, Zapico SC, Van De Voorde W, Decorte R. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. *Epigenetics.* 2015;5:1922–30.
56. Choi H, Joe S, Nam H. Development of tissue-specific age predictors using DNA methylation data. *Genes (Basel).* 2019;10:888.
57. Xu C, et al. A novel strategy for forensic age prediction by DNA methylation and support vector regression model. *Sci Reports.* 2015;5:17788.
58. Boroni M, et al. Highly accurate skin-specific methylome analysis algorithm as a platform to screen and validate therapeutics for healthy aging. *Clin Epigenetics.* 2020;12:1–16.
59. Everson TM, et al. DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. *Genome Med.* 2015;7:89.
60. Baur B, Bozdag S. A feature selection algorithm to compute gene centric methylation from probe level methylation data. *PLoS ONE.* 2016;11:e0148977.
61. Sugden K, et al. Patterns of reliability: assessing the reproducibility and integrity of DNA methylation measurement. *Patterns.* 2020;1: 100014.
62. Logue MW, et al. The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics.* 2017;9:1363–71.
63. Taylor J, Tibshirani R. Post-selection inference for -penalized likelihood models. *Can J Stat.* 2018;46:41–61.
64. de Lima Camillo LP, Lapierre LR, Singh R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging.* 2022;8(1):4.
65. McCartney DL, et al. An epigenome-wide association study of sex-specific chronological ageing. *Genome Med.* 2019;12:1–11.
66. Thymus Regeneration, Immunorestitution, and Insulin Mitigation Extension Trial - Full Text View - ClinicalTrials.gov. Available at: <https://clinicaltrials.gov/ct2/show/NCT04375657>. Accessed 4 July 2022

67. Lu Y, et al. Reprogramming to recover youthful epigenetic information and restore vision. *Nat.* 2020;588:124–9.
68. Olova N, Simpson DJ, Marioni RE, Chandra T. Partial reprogramming induces a steady decline in epigenetic age before loss of somatic identity. *Aging Cell.* 2019;18:e12877.
69. Strandberg TE, Tilvis RS. C-reactive protein, cardiovascular risk factors, and mortality in a prospective study in the elderly. *Arterioscler Thromb Vasc Biol.* 2000;20:1057–60.
70. Lobo SMA, et al. C-reactive protein levels correlate with mortality and organ failure in critically ill patients. *Chest.* 2003;123:2043–9.
71. Mendall MA, et al. C-reactive protein: relation to total mortality, cardiovascular mortality and cardiovascular risk factors in men. *Eur Heart J.* 2000;21:1584–90.
72. Furman D, et al. Chronic inflammation in the etiology of disease across the life span. *Nat Med.* 2019;25(12):1822–32.
73. Chung HY, et al. Redefining chronic inflammation in aging and age-related diseases: proposal of the senoinflammation concept. *Aging Dis.* 2019;10:367.
74. Gesing A, et al. A long-lived mouse lacking both growth hormone and growth hormone receptor: a new animal model for aging studies. *J Gerontol A Biol Sci Med Sci.* 2017;72:1054–61.
75. Junnila RK, et al. Disruption of the GH receptor gene in adult mice increases maximal lifespan in females. *Endocrinology.* 2016;157:4502–13.
76. Bartke A. Growth hormone and aging: updated review. *World J Mens Health.* 2019;37:19.
77. Aguiar-Oliveira MH, Bartke A. Growth hormone deficiency: health and longevity. *Endocr Rev.* 2019;40:575–601.
78. Laron Z. Do deficiencies in growth hormone and insulin-like growth factor-1 (IGF-1) shorten or prolong longevity? *Mech Ageing Dev.* 2005;126:305–7.
79. Gao X, Jia M, Zhang Y, Breitling LP, Brenner H. DNA methylation changes of whole blood cells in response to active smoking exposure in adults: a systematic review of DNA methylation studies. *Clin Epigenetics.* 2015;7:1–10.
80. Zhang Y, et al. F2RL3 methylation in blood DNA is a strong predictor of mortality. *Int J Epidemiol.* 2014;43:1215–25.
81. Zhang Y, et al. DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat Commun.* 2017;8:14617.
82. Ono M, et al. Foxp3 controls regulatory T-cell function by interacting with AML1/Runx1. *Nature.* 2007;446:685–9.
83. Fu L, et al. High expression of ETS2 predicts poor prognosis in acute myeloid leukemia and may guide treatment decisions. *J Transl Med.* 2017;15:1–9.
84. Liquori A, et al. Acute promyelocytic leukemia: a constellation of molecular events around a single PML-RARA fusion gene. *Cancers (Basel).* 2020;12:624.
85. Bernabeu, E. cAge/bAge Github Repository. GitHub (2022). Available at: [https://github.com/elenabernabeu/cage\\_bage](https://github.com/elenabernabeu/cage_bage).
86. Bernabeu, E. et al. cAge and bAge EWAS - Epigenome-wide association study of chronological age (linear and quadratic) and of survival, . University of Edinburgh. College of Medicine & Veterinary Medicine. Institute of Genetics and Cancer. Centre for Genomic & Experimental Medicine (CGEM) (2022). <https://doi.org/10.7488/ds/3792>.

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

