

Refining epigenetic prediction of chronological and biological age

Elena Bernabeu¹, Daniel L McCartney¹, Danni A Gadd¹, Robert F Hillary¹, Ake T Lu^{2,3}, Lee Murphy⁴, Nicola Wrobel⁴, Archie Campbell¹, Sarah E Harris⁵, David Liewald⁵, Caroline Hayward^{1,6}, Cathie Sudlow^{7,8,9}, Simon R Cox⁵, Kathryn L Evans¹, Steve Horvath^{2,3}, Andrew M McIntosh^{1,10}, Matthew R Robinson¹¹, Catalina A Vallejos^{6,12}, Riccardo E Marioni^{1*}

¹ Centre for Genomic and Experimental Medicine, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

² Department of Human Genetics, David Geffen School of Medicine, University of California, Los Angeles, CA, USA

³ Altos Labs, San Diego, USA

⁴ Edinburgh Clinical Research Facility, University of Edinburgh, Edinburgh, UK

⁵ Lothian Birth Cohorts, Department of Psychology, University of Edinburgh, Edinburgh, UK

⁶ Medical Research Council Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK

⁷ Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, UK

⁸ BHF Data Science Centre, Health Data Research UK, London, UK

⁹ Usher Institute, Edinburgh Medical School, University of Edinburgh, Edinburgh, UK

¹⁰ Division of Psychiatry, University of Edinburgh, Royal Edinburgh Hospital, Edinburgh, UK

¹¹ Institute of Science and Technology Austria, Klosterneuburg, Austria

¹² The Alan Turing Institute, London, UK

* Correspondence: Riccardo E Marioni, riccardo.marioni@ed.ac.uk

1 **Abstract**

2

3 Epigenetic clocks can track both chronological age (cAge) and biological age (bAge). The latter is
4 typically defined by physiological biomarkers and risk of adverse health outcomes, including all-cause
5 mortality. As cohort sample sizes increase, estimates of cAge and bAge become more precise. Here,
6 we aim to refine predictors and improve understanding of the epigenomic architecture of cAge and
7 bAge. First, we perform large-scale ($N = 18,413$) epigenome-wide association studies (EWAS) of
8 chronological age and all-cause mortality. Next, to improve cAge prediction, we use methylation data
9 from 24,673 participants from the Generation Scotland (GS) study, the Lothian Birth Cohorts (LBC) of
10 1921 and 1936 and 8 publicly available datasets. Through the inclusion of linear and non-linear age-
11 CpG associations from the EWAS, feature pre-selection/dimensionality reduction in advance of elastic
12 net regression, and a leave-one-cohort-out (LOCO) cross validation framework, we arrive at an
13 improved cAge predictor (median absolute error = 2.3 years across 10 cohorts). In addition, we train a
14 predictor of bAge on 1,214 all-cause mortality events in GS, based on epigenetic surrogates for 109
15 plasma proteins and the 8 component parts of GrimAge, the current best epigenetic predictor of all-
16 cause mortality. We test this predictor in four external cohorts (LBC1921, LBC1936, the Framingham
17 Heart Study and the Women's Health Initiative study) where it outperforms GrimAge in its association
18 to survival ($HR_{GrimAge} = 1.47$ [1.40, 1.54] with $p = 1.08 \times 10^{-52}$, and $HR_{bAge} = 1.52$ [1.44, 1.59] with $p =$
19 2.20×10^{-60}). Finally, we introduce MethylBrowsR, an online tool to visualize epigenome-wide CpG-age
20 associations.

21

22 **Introduction**

23

24 The development and application of epigenetic predictors for healthcare research has grown
25 dramatically over the last decade¹. These predictors can aid disease risk stratification, and are based
26 on associations between CpG DNA methylation (DNAm) and age, health, and lifestyle outcomes.
27 DNAm is dynamic, tissue-specific and is influenced by both genetic and environmental factors. DNAm
28 can precisely track ageing through predictors termed "epigenetic clocks"²⁻⁸. DNAm scores have also
29 been found to capture other components of health, such as smoking status^{9,10}, alcohol consumption^{11,12},
30 obesity^{11,13}, and protein levels¹⁴.

31
32 “First generation” epigenetic ageing clocks, including those by Horvath³ and Hannum et al⁴, were trained
33 on chronological age²⁻⁴ (cAge), with near-perfect clocks expected to arise as sample sizes grow⁵.
34 However, cAge clocks hold limited capability for tracking and quantifying age-related health status, also
35 termed biological age (bAge)^{5,8}. To address this, “second generation” clocks have been trained on other
36 age-related measures, including a phenotypic biomarker of morbidity (PhenoAge¹⁵), rate of ageing
37 (DunedinPoAm¹⁶), and time to all-cause mortality (GrimAge¹⁷). Regressing an epigenetic clock predictor
38 (whether trained on cAge or bAge) on chronological age within a cohort gives rise to an “age
39 acceleration” residual with positive values corresponding to faster biological ageing.
40
41 Penalised regression approaches such as elastic net¹⁸ are used to derive epigenetic predictors. Such
42 epigenetic clocks typically capture a weighted linear combination of CpGs that optimally predict an
43 outcome from a statistical perspective i.e. no preference is given to the location or possible biological
44 role of the input features. The majority consider genome-wide CpG sites as potential predictive features.
45 However, others have used a two-stage approach that first creates DNAm surrogates (or epigenetic
46 scores - EpiScores) for biomarkers (also typically via elastic net) prior to training a second elastic net
47 model on the phenotypic outcome or time-to-event^{14,17}. GrimAge is currently the gold-standard bAge
48 epigenetic clock. It is derived from age, sex, and EpiScores of smoking pack years and 7 plasma
49 proteins that have been associated with mortality or morbidity: adrenomedullin (ADM), beta-2-
50 microglobulin (B2M), cystatin C, growth differentiation factor 15 (GDF15), leptin, plasminogen activation
51 inhibitor 1 (PAI1), and tissue inhibitor metalloproteinase (TIMP1). Recently, a wider set of 109
52 EpiScores for the circulating proteome were generated by Gadd et al¹⁴. These have not yet been
53 considered as potential features for the prediction of bAge.
54
55 Here, we improve the prediction of both cAge and bAge (**Figure 1**). We first present large-scale
56 epigenome-wide association studies (EWAS) of age (for both linear and quadratic CpG effects) and all-
57 cause mortality. A predictor of cAge is then generated using DNAm data from 13 cohorts, including
58 samples from >18,000 participants of the Generation Scotland study¹⁹. We use a leave-one-cohort-out
59 (LOCO) prediction framework, including dimensionality reduction prior to feature selection for linear and
60 non-linear DNAm-age relationships (ascertained through the EWAS), and test it on ten external

61 datasets. Through data linkage to death records, we develop a bAge predictor of all-cause mortality,
62 which we compare against the current gold-standard predictor, GrimAge, in four external cohorts. These
63 analyses highlight the potential for large DNAm resources to generate increasingly accurate predictors
64 of (i) cAge, with potential forensic utility, and (ii) bAge, with potential implications for risk prediction and
65 clinical trials.

66

67 **Results**

68

69 Data overview

70

71 Generation Scotland is a Scottish family-based study with over 24,000 participants recruited between
72 2006 and 2011¹⁹. Blood-based DNAm levels at 752,722 CpG sites were quantified using the Illumina
73 MethylationEPIC array for 18,413 individuals (see **Methods**). Participants were aged between 18 and
74 99 years at recruitment, with a mean age of 47.5 years (SD 14.9, **Table 1**). A total of 1,214 participant
75 deaths have been recorded as of March 2022, via linkage to the National Health Service Central
76 Register, provided by the National Records of Scotland.

77

78 In order to train and test a cAge predictor, data from an additional 6,260 individuals from ten external
79 cohorts were considered. These included the Lothian Birth Cohorts (LBC) of 1921 and 1936, and eight
80 publicly available Gene Expression Omnibus (GEO) datasets (see Methods, **Table 1**). Given that the
81 external datasets assessed DNAm (blood-based apart from GSE78874, which considered saliva) using
82 the Illumina HumanMethylation450K array, the Generation Scotland data were subset to 374,791 CpGs
83 that were present across all studies.

84

85 To test the bAge predictor, data from an additional 4,134 individuals (with a total of 1,653 deaths) from
86 six external cohorts were considered. These included both the LBC1921 and LBC1936 cohorts, as well
87 as the Framingham Heart Study (FHS) and the Women's Health Initiative (WHI) Broad Agency Award
88 23 (B23) study for Black, White, and Hispanic individuals (see Methods, **Table 2**).

89

90 Epigenome-wide association studies of cAge

91

92 EWAS of cAge were performed in the Generation Scotland cohort, resulting in 99,832 linear and

93 137,195 quadratic CpG associations that were epigenome-wide significant ($p < 3.6 \times 10^{-8}$,

94 **Supplementary Figure 1, Supplementary Table 1 and 2**, see **Methods**). These mapped to 17,339

95 and 19,432 unique genes, respectively. There were 48,312 CpGs with both a significant linear and

96 quadratic association.

97

98 The most significant linear associations included cg16867657 and cg24724428 (*ELOVL2*), cg08097417

99 (*KLF14*), and cg12841266 (*LHFPL4*), all $p < 1 \times 10^{-300}$, (**Supplementary Table 1, Supplementary**

100 **Figure 2**). Around half of the CpGs with a significant linear association (51,213/99,832, 51.3%) showed

101 a positive association with age. The most significant quadratic associations were cg11084334 (*LHFPL4*,

102 $p = 6.49 \times 10^{-206}$), cg15996534 (*LOC134466*, $p = 8.7 \times 10^{-194}$), and cg23527621 (*ECE2* and *CAMK2N2*,

103 $p = 9.95 \times 10^{-189}$, **Supplementary Table 2, Supplementary Figure 3**).

104

105 The univariate associations between all 752,722 CpGs and cAge in a subset of 4,450 unrelated

106 participants (DNAm arrays processed together in a single experiment) from Generation Scotland can

107 be visualised via an online ShinyApp, MethylBrowsR (<https://shiny.igmm.ed.ac.uk/MethylBrowsR/>).

108

109 Prediction of cAge

110

111 Epigenetic clocks for cAge were created using elastic net penalised regression. Input features consisted

112 of CpG and CpG² DNAm values for sites that were epigenome-wide significant in their corresponding

113 EWAS analysis (see **Methods, Figure 2**). After iterating through combinations of CpG and CpG² terms

114 (ranked by EWAS p-value), the best-performing model considered the top 10,000 CpG and top 300

115 CpG² sites from the EWAS as potentially informative features (see **Methods, Supplementary Table 3**

116 **and 4, Supplementary Figure 4 and 5**). A single external cohort was used for this screening step

117 (GSE40279, N = 656) and model fit was based on the root mean squared error (RMSE) and median

118 absolute error (MAE) of prediction.

119

120

121 A LOCO framework was used to train the cAge predictor, whereby for each of the 10 external cohorts,
122 a model was trained on data from Generation Scotland and the remaining nine external cohorts. Testing
123 was then performed on the excluded cohort (total $N_{\text{testing}} = 6,260$). A final model was also trained on all
124 11 cohorts ($N_{\text{training}} = 24,673$).

125

126 Both age and log(age) were considered as outcomes, with the latter showing better prediction results
127 in younger individuals, reflecting the importance of considering non-linear DNAm-age associations in
128 cAge prediction. As a result, if the initial cAge prediction was <20 years, that individual's predicted age
129 was re-estimated using weights from the log(age) model.

130

131 The combined LOCO prediction results showed a strong correlation with cAge ($r = 0.96$, **Figure 3, Table**
132 **1**) and a MAE of 2.3 years. Furthermore, 24% of individuals were classified to within one year of their
133 chronological age. The cohort with the largest prediction errors was GSE78874, in which DNAm was
134 measured in saliva instead of blood.

135

136 The elastic net model (trained in all 11 cohorts) with the lowest mean cross-validated error identified
137 2,330 features (2,274 linear and 56 quadratic) as most predictive of age, and 1,986 features (1,931
138 linear and 55 quadratic) as most predictive of log(age). The weights for the age model are presented in
139 **Supplementary Table 5**, and for the log(age) model in **Supplementary Table 6**.

140

141 Epigenome-wide association study of all-cause mortality

142

143 To identify individual CpG loci associated with survival, an EWAS on time to all-cause mortality was
144 performed in Generation Scotland ($N_{\text{deaths}} = 1,214$, see **Methods**). This analysis identified 1,182
145 epigenome-wide significant associations ($p < 3.6 \times 10^{-8}$, **Supplementary Figure 6**), which mapped to
146 704 unique genes. Around a third (418/1,182 = 35.36%) of these CpGs were associated with a
147 decreased survival time. The lead findings included CpGs mapping to smoking-related loci^{10,20-24} such
148 as cg05575921 (*AHRR*, $p = 3.01 \times 10^{-57}$), cg03636183 (*F2RL3*, $p = 6.78 \times 10^{-44}$), cg19859270 (*GPR15*,
149 $p = 1.09 \times 10^{-33}$), cg17739917 (*RARA*, $p = 1.92 \times 10^{-33}$), cg14391737 (*PRSS23*, $p = 5.59 \times 10^{-33}$),
150 cg09935388 (*GFI1*, $p = 3.30 \times 10^{-31}$), and cg25845814 (*ELMSAN1/MIR4505*, $p = 1.31 \times 10^{-30}$)

151 (Supplementary Table 7). Of the non-smoking-related CpGs amongst the top 50 associations, seven
152 mapped to genes whose methylation has been linked to various forms of cancer, including *ZMIZ1*²⁵,
153 *SOCS3*²⁶⁻²⁸, *ZMYND8*²⁹ and *CHD5*³⁰⁻³². Another probe mapped to *FKBP5*, a gene whose methylation
154 is involved in the regulation of the stress response, and which has been linked to increased
155 cardiometabolic risk through accelerated ageing³³. Finally, one top probe mapped to *SKI*, whose
156 methylation has been linked to age-related macular degeneration³⁴. All associations remained after
157 adjusting for relatedness in the Generation Scotland cohort (see Methods, Supplementary Table 8).

158

159 There was a high correlation of the Z-score effect sizes across the 200 sites that overlapped between
160 our study and the 257 epigenome-wide significant findings from a recent large (N = 12,300, N_{deaths} =
161 2,561) meta-analysis of all-cause mortality ($r = 0.58$, Supplementary Figure 7). All 200 sites were
162 significant at a nominal $p < 0.05$ threshold and 25 were epigenome-wide significant at $p < 3.6 \times 10^{-8}$.

163

164 A gene-set enrichment analysis considering genes to which epigenome-wide significant CpGs mapped
165 to returned 198 significantly enriched (FDR $p < 0.05$) GO biological processes (see Methods, full FUMA
166 gene-set enrichment results in Supplementary Table 9). The most significantly enriched GO terms
167 included processes relating to neurogenesis/neuron differentiation and development, positive immune
168 system regulation and development, cell motility and organization, and regulation of protein
169 modification/phosphorylation. Other significantly enriched sets included sites bound by FOXP3, ETS2,
170 and the PML-RARA fusion protein.

171

172 Prediction of bAge

173

174 Amongst the second generation epigenetic clocks, GrimAge is the current best predictor of lifespan
175 (time to death)¹⁷. In an effort to improve the prediction of bAge, an elastic net Cox model was trained
176 on all-cause mortality in Generation Scotland (N_{total} = 18,365, N_{deaths} = 1,214, see Methods). The
177 GrimAge components (age, sex, and EpiScores for smoking and 7 plasma proteins) and Gadd et al's
178 109 protein EpiScores¹⁴ were considered as potentially-informative features (Figure 4).

179

180 The elastic net Cox model identified a weighted sum of 35 features as most predictive of all-cause
181 mortality in Generation Scotland. These included age and the GrimAge smoking EpiScore, along with
182 5/7 protein EpiScores from GrimAge (B2M, cystatin C, GDF15, PAI1, and TIMP1), and 28/109 protein
183 EpiScores from Gadd et al¹⁴. Amongst these were EpiScores for C-reactive protein (CRP), the growth
184 hormone receptor (GHR) protein, and numerous cytokines (CCL11, CCL23, CCL18, CXCL10, CXCL9,
185 CXCL11, and HGF). The weights for the linear predictor are presented in **Supplementary Table 10**.

186

187 The bAge predictor was regressed on age to obtain a measure of epigenetic age acceleration
188 (bAgeAccel). The epigenetic age acceleration residuals showed significant associations with all-cause
189 mortality across four test cohorts of differing ancestries (**Table 2, Supplementary Table 11, Figure 5**).
190 The bAge measure showed slightly stronger associations than GrimAge (also regressed on age, termed
191 GrimAgeAccel) in fixed effects meta-analyses (Hazard Ratio and 95% Confidence Interval per SD
192 difference of GrimAgeAccel and bAgeAccel: HR = 1.47 [1.40, 1.54] with $p = 1.08 \times 10^{-52}$, and HR = 1.52
193 [1.44, 1.59] with $p = 2.20 \times 10^{-60}$, respectively.

194

195 **Discussion**

196

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

200

201 Epigenetic cAge prediction is expected to reach near-perfect estimates as sample sizes grow⁵. Making
202 use of Generation Scotland, a very large single-cohort DNAm resource, we derived a cAge predictor
203 with a MAE of 2.3 years, tested in over 6,000 external samples. Our predictor has potential forensic
204 applications, although ethical caveats exist⁸. In addition, despite the high correlations and low RMSE
205 and MAE estimates at the population level, there are still several individuals with inaccurate predictions
206 (e.g. > 20 years between predicted and actual age, **Figure 3**), though this could also reflect sample
207 mix-ups or data entry errors.

208

209 cAge prediction was improved when accounting for non-linear relationships between DNAm and age.
210 Whilst generally understudied, non-linear patterns have been found at numerous CpG sites, where
211 DNAm is found to increase rapidly in early ages and stabilize in adulthood, potentially reflecting
212 developmental processes³⁵. Similarly, stable DNAm levels followed by rapid methylation/demethylation
213 have also been described in later life³⁶, which could offer insight into aging-specific processes. Given
214 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
216 population. Future studies could also consider sex-specific models as diverging non-linear patterns
217 between males and females have been shown in previous studies³⁷. Interactions between CpGs along
218 with higher order polynomial terms and spline-based models might better capture some of these non-
219 linear changes.

220
221 The development of the cAge predictor highlighted the advantages of feature pre-selection ahead of
222 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, pre-selection greatly improved performance ($r = 0.96$, RMSE = 3.92, MAE = 2.32). 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 net³⁸⁻⁴⁵. Whereas the feature pre-selection here required arbitrary decisions on thresholds, other studies have found that feature reduction via PCA optimises DNAm predictors^{46,47}.

228
229 Feature pre-selection may have aided cAge predictions by screening out CpGs with low intra-sample
230 variability due to technical variance^{48,49}. One previous study⁴⁷ observed that CpGs with stronger cAge
231 associations were more reliable. A limitation of our approach to feature pre-selection was that it was
232 biased towards the Generation Scotland cohort in which the age EWAS were conducted. We also note
233 that pre-selection introduces statistical challenges associated to post-selection inference⁵⁰.
234 Furthermore, our penalised regression modelling strategy for cAge only incorporated additive effects.
235 Non-additive tree ensemble methods and other machine learning frameworks may improve predictions
236 further⁵¹. Finally, as our predictor has been mainly trained and tested on blood data, it may not
237 generalise to other tissues.

238

239 Whilst a single DNAm predictor of cAge is of interest, the selected CpG features are unlikely to identify
240 all epigenome-wide patterns related to ageing. Our EWAS of chronological age identified 99,832 linear
241 and 137,195 quadratic CpG-age associations. The sample size was more than double that of the largest
242 study reported on the EWAS Catalog⁵² - our previous Generation Scotland analysis⁵³. In addition to
243 refining our previously described DNAm-age linear associations, we have extended previous small-
244 scale approaches to highlight non-linear patterns^{36,37}. As shown here, these findings can aid the
245 predictive performance of epigenetic clocks, and may additionally improve our understanding of
246 epigenetic changes during development and ageing-related decline in later life.

247

248 Recent work has shifted focus from the prediction of cAge to bAge, with more expansive clinical
249 applications. Our new bAge predictor of all-cause mortality had a greater effect size and was more
250 statistically significant than GrimAge in the external test set meta-analysis. GrimAge is already being
251 used as an end-point for clinical trials⁵⁴ and studies of rejuvenation^{55,56}. The bAge predictor included
252 EpiScores for CRP and numerous cytokines, which reflect inflammation and predict overall and
253 cardiovascular mortality⁵⁷⁻⁵⁹. Chronic inflammation can lead to several diseases, including
254 cardiovascular disease and exacerbates the ageing process^{60,61}. In addition, the growth hormone
255 receptor (GHR) protein EpiScore was selected; both the receptor and its corresponding protein have
256 been linked to longevity in mouse models⁶²⁻⁶⁶. 25/28 of the selected EpiScores from Gadd et al¹⁴ have
257 been associated to multiple diseases, including diabetes, chronic obstructive pulmonary disease,
258 ischaemic heart disease, lung cancer, Alzheimer's, rheumatoid arthritis, stroke, and depression
259 (**Supplementary Table 10**). As sample sizes for cause-specific mortality outcomes increase, a more
260 granular suite of lifespan predictors can be developed.

261

262 Whereas the cAge predictions translated into external cohorts with minimal calibration issues,
263 individual-level bAge predictions were highly variable. Future work for these (and all) DNAm array-
264 based predictors should consider the limitations of signatures that lack absolute thresholds/cut-points
265 for risk prediction in a new individual selected at random from the population.

266

267 A total of 1,182 epigenome-wide significant associations were identified in our EWAS of all-cause
268 mortality. The most significant probes mapped to genes previously associated with smoking, such as

269 *AHRR*, *F2RL3*, and *GPR15*⁶⁷. Hypomethylation at probes nearby these genes has been previously
270 linked to increased mortality risk, be that all-cause or disease specific (e.g., cancer or, cardiovascular-
271 related mortality)^{20,68-70}. Other, non-smoking related, lead probes mapped to genes whose methylation
272 has been linked to various forms of cancer, increased cardiometabolic risk, and age-related macular
273 degeneration²⁵⁻³⁴. There was moderate agreement (correlation of 0.58 between Z scores) between our
274 findings and the significant results from a previous EWAS meta-analysis of survival. However, different
275 covariates and ancestries were considered across these studies. An enrichment analysis highlighted
276 links to neurodevelopment and immune regulation, as well as to sites bound by *FOXP3*, *ETS2*, and the
277 *PML-RARA* fusion protein. *FOXP3* is a transcriptional regulator involved in the development and
278 inhibitory function of regulatory T-cells⁷¹. *ETS2* and *PML-RARA* are a protooncogene and a protein
279 resulting from a chromosomal translocation that resulting in an oncofusion protein, respectively, having
280 both been linked to acute myeloid leukemia^{72,73}. This finding may be influenced by the large number of
281 cancer-related deaths in Generation Scotland (N = 509). Further work is needed to disentangle the role
282 of methylation/demethylation at these sites with survival. Future EWAS on specific mortality causes will
283 highlight mechanisms underlying age- and disease-related decline.

284

285 The integration of multiple large datasets and new approaches to feature selection has facilitated
286 improvements to the blood-based epigenetic prediction of biological and chronological age. The
287 inclusion of multiple protein EpiScore features and consideration of quadratic DNA effects may also
288 be relevant for other EWAS and prediction studies. Together, this can improve our biological
289 understanding of complex traits and the prediction of adverse health outcomes.

290

291 **Methods**

292

293 Generation Scotland

294

295 *Cohort description*

296

297 Generation Scotland: Scottish Family Health Study is a population-based cohort study that includes
298 ~7,000 families from across Scotland¹⁹. Study recruitment took place between 2006 and 2011 when

299 participants were aged between 17 and 99 years (**Table 1**). In addition to completing health and lifestyle
300 questionnaires, participants donated blood or saliva samples for biomarker and omics analyses. The
301 majority of participants also provided consent for linkage to their electronic medical records, yielding
302 retrospective and prospective information on primary and secondary disease outcomes as well as
303 prescription data.

304

305 *Data linkage to death records*

306

307 Information on mortality and cause of death is routinely updated via linkage to the National Health
308 Service Central Register, provided by the National Records of Scotland. The data used here were
309 correct as of March 2022, with a total of 1,214 deaths and 18,365/18,413 samples with non-missing
310 and non-negative time-to-death/event (TTE) values. Average TTE amongst deaths was 7.79 (SD 3.54)
311 years. Leading causes of death included malignant neoplasms (509), ischaemic heart disease (134),
312 cerebrovascular disease (69), other forms of heart disease (44), chronic lower respiratory disease (42),
313 mental disorders including dementia (36), and other degenerative diseases of the nervous system (35).

314

315 *DNA methylation in Generation Scotland*

316

317 DNA methylation in blood was quantified for 18,413 Generation Scotland participants across three
318 separate sets ($N_{Set1} = 5,087$, $N_{Set2} = 4,450$, $N_{Set3} = 8,876$) using the Illumina MethylationEPIC (850K)
319 array. Individuals in Set 1 included a mixture of related and unrelated individuals. Set 2 comprised
320 individuals unrelated to each other and also to those in Set 1. Set 3 contained a mix of related individuals
321 – both to each other and to those in Sets 1 and 2 – and included all remaining samples available for
322 analysis.

323

324 Quality control details have been reported previously^{53,74}. Briefly, probes were removed based on (i)
325 outliers from visual inspection of the log median intensity of the methylated versus unmethylated signal
326 per array, (ii) a bead count < 3 in more than 5% of samples, (iii) ≥ 5% of samples having a detection *p*-
327 value > 0.05, (iv) if they pertained to the sex chromosomes, (v) if they overlapped with SNPs, and/or
328 (vi) if present in potential cross-hybridizing locations⁷⁵. Samples were removed (i) if there was a

329 mismatch between their predicted sex and recorded sex, (ii) if $\geq 1\%$ of CpGs had a detection *p*-value
330 > 0.05 , (iii) if sample was not blood-based, and/or (iv) if participant responded “yes” to all self-reported
331 diseases in questionnaires. Dasen normalisation⁷⁶ was carried out per set (for cAge training) or across
332 all individuals (for EWAS). A total of 752,722 CpGs remained after QC. To maximise the generalisability
333 of the predictors across different versions of Illumina arrays, we subset the content to the intersection
334 of sites on the EPIC and 450k arrays, as well as to those present across all cohorts considered in the
335 study (**Table 1**), totalling 374,791 CpGs.

336

337 External datasets

338

339 To test the cAge predictor, we considered DNA methylation for a total of 6,260 external samples, from
340 eight publicly available datasets from the Gene Expression Omnibus (GEO) resource and repeated
341 measures (up to four time points) from two cohorts of blood-based DNAm, the Lothian Birth Cohorts
342 (LBC) of 1936 and 1921 (**Table 1**)^{4,77–82}. The baseline samples from the LBC cohorts, along with the
343 Framingham Heart Study (FHS) and the Women’s Health Initiative (WHI) study, were also used for the
344 testing of our bAge predictor (**Table 2**).

345

346 *Lothian Birth Cohorts*

347

348 LBC1921 and LBC1936 are longitudinal studies of ageing on individuals born in 1921 and 1936,
349 respectively⁷⁷. Study participants completed the Scottish Mental Surveys of 1932 and 1947 at
350 approximately age 11 years old and were living in the Lothian area of Scotland at the time of recruitment
351 in later life. Blood samples considered here were collected at around age 79 for LBC1921, and at around
352 age 70 for LBC1936. DNA methylation was quantified using the Illumina HumanMethylation450 array,
353 for a total of 692 (up to 3 repeated measurements from 469 individuals) and 2,795 (up to 4 repeated
354 measurements from 1,043 individuals) samples from LBC1921 and LBC1936 respectively. Quality
355 control details have been reported previously^{5,83}. Briefly, probes were removed (i) if they presented a
356 low (< 95%) detection rate with *p*-value < 0.01 , and/or (ii) if they presented inadequate hybridization,
357 bisulfite conversion, nucleotide extension, or staining signal, as assessed by manual inspection.
358 Samples were removed (i) if they presented a low call rate (<450,000 probes detected at *p*-value $<$

359 0.01) and/or (ii) if predicted sex did not match reported sex. Finally, as stated previously, probes were
360 filtered down to the 374,791 common across all datasets (**Table 1**). Missing values were mean imputed.

361
362 A total of 421 and 895 samples from LBC1921 and LBC1936 respectively, corresponding to the first
363 wave of each study (thus aged around 79 and 70 at time of sampling for each cohort respectively), were
364 used in our bAge analysis (**Table 2**). All-cause mortality was assessed via linkage to the National Health
365 Service Central Register, provided by the National Records of Scotland. The data used here are correct
366 as of January, 2022, with a total of 421 and 367 deaths in LBC1921 and LBC1936 respectively.

367
368 *Gene Expression Omnibus (GEO) datasets*

369
370 DNAm and age information for 2,773 individuals from a total of 8 datasets was downloaded from the
371 public domain (Gene Expression Omnibus, GEO). DNAm was quantified with Illumina's
372 HumanMethylation450 chip. QC information can be found in each pertaining publication (**Table 1**), and
373 CpGs were filtered down to the 374,791 common across all datasets. Missing values were mean
374 imputed.

375
376 *Framingham Heart Study (FHS)*

377
378 The FHS cohort is a large-scale longitudinal study started in 1948, initially investigating the common
379 factors of characteristics that contribute to cardiovascular disease (CVD)⁸⁴. The study at first enrolled
380 participants living in the town of Framingham, Massachusetts, who were free of overt symptoms of
381 CVD, heart attack or stroke at enrolment. In 1971, the study established the FHS Offspring Cohort to
382 enrol a second generation of the original participants' adult children and their spouses for conducting
383 similar examinations⁸⁵. Participants from the FHS Offspring Cohort were eligible for our study if they
384 attended both the seventh and eighth examination cycles and consented to having their molecular data
385 used for study. We used data pertaining to a total of 711 individuals which had not been used in the
386 training of GrimAge, and for which DNAm data and death records were available. Peripheral blood
387 samples were obtained on the eighth examination cycle, and DNAm data was measured using the

388 Illumina Infinium HumanMethylation450 array, with QC details are described elsewhere¹⁷. Deaths
389 recorded are accurate as of 1st January 2013, with a total of 100 recorded.

390

391 *Women's Health Initiative (WHI)*

392

393 The WHI study enrolled postmenopausal women aged 50-79 years into the clinical trials (CT) or
394 observational study (OS) cohorts between 1993 and 1998. We included 2,107 women from "Broad
395 Agency Award 23" (WHI BA23). WHI BA23 focuses on identifying miRNA and genomic biomarkers of
396 coronary heart disease (CHD), integrating the biomarkers into diagnostic and prognostic predictors of
397 CHD and other related phenotypes. This cohort is divided into three datasets, pertaining to three
398 different ancestries: White, Black, and Hispanic, with 998, 676, and 433 participants respectively. Blood-
399 derived DNAm data was available for participants. DNAm data was measured using the Illumina
400 Infinium HumanMethylation450 array, QC details described elsewhere¹⁷. Deaths recorded are accurate
401 as March 1st, 2017, with a total of 418, 229, and 118 recorded for White, Black, and Hispanic ancestries
402 respectively.

403

404 EWAS of chronological age

405

406 We conducted an EWAS to identify CpG sites that had linear or quadratic associations with
407 chronological age, using Generation Scotland data (N = 18,413, CpGs = 752,722). Linear regression
408 analyses were carried out which included both linear and quadratic CpG M-values as predictor variables
409 and age as the dependent variable (Age ~ CpG and Age ~ CpG + CpG², respectively). Fixed effect
410 covariates included estimated white blood cell (WBC) proportions (basophils, eosinophils, natural killer
411 cells, monocytes, CD4T, and CD8T cells) calculated in the *minfi* R package (version 1.36.0)⁸⁶ using the
412 Houseman method⁸⁷, sex, DNAm batch/set, smoking status (current, gave up in the last year, gave up
413 more than a year ago, never, or unknown), smoking pack years, and 20 methylation based principal
414 components (PCs) to correct for unmeasured confounders. Age was centered by its mean, and CpG
415 and CpG² M-values were scaled to mean zero and variance one. Epigenome-wide significance was set
416 at *p*-value < 3.6 x 10⁻⁸, as per Saffari et al⁸⁸.

417

418 Prediction of chronological age

419

420 Elastic net regression (with $\alpha = 0.5$ as the L_1 , L_2 mixing parameter) was used to derive a predictor of
421 chronological age from the 374,791 CpG sites common across all cohorts considered in cAge training
422 (description of cohorts in **Table 1**). The *biglasso* R package (version 1.5.1) was used⁸⁹, with 25-fold
423 cross validation (CV) to select the shrinkage parameter (λ) that minimised the mean cross-validated
424 error. This resulted in randomly assigned folds of ~1,000 individuals. A sensitivity analysis was
425 performed, assigning individuals from the same methylation set and cohort to individual folds, which
426 returned highly similar results.

427

428 *Leave-one-cohort-out (LOCO)*

429

430 The cAge predictor was created and tested using a leave-one-cohort-out (LOCO) framework, where
431 the model was trained in 10 cohorts and tested on the excluded external cohort (**Figure 2**). The final
432 reported model was trained using all 11 sets described here. Pearson correlations (r) with reported age
433 were calculated along with the root mean square error (RMSE) and median absolute error (MAE).

434

435 *Log(age)*

436

437 In addition to training on chronological age, models were also trained on the natural logarithm of
438 chronological age, *log(age)*. The age of our test samples was predicted using the model fit on
439 chronological age, and, if the predicted age returned was 20 years or younger, a new prediction was
440 obtained making use of the model fit on *log(age)*. This approach parallels that in Horvath's 2013 clock,
441 which log-transforms chronological age in under 20s prior to training³.

442

443 *Feature pre-selection*

444

445 Several studies have highlighted the benefits of feature pre-selection for elastic net^{46,47}. Here, we
446 performed preliminary analyses, including differently sized subsets of CpG sites as features in elastic
447 net. We considered sites that were epigenome-wide significant at $p < 3.6 \times 10^{-8}$ and then ranked CpGs

448 in ascending order of *p*-value (most significant ranked first), before defining subsets of varying sizes
449 (from 1,000 to 300,000 CpGs). Our training cohort was Generation Scotland, whilst our test set was
450 GSE40279, one of the largest external datasets with the widest age range. Our analyses showed that
451 the 10,000 most significant loci (age - CpG associations) yielded the test set predictions with the highest
452 *r* and lowest RMSE (**Supplementary Table 3, Supplementary Figure 4**). In addition to these sites,
453 subsets of CpGs with a significant quadratic relationship to age were explored, with subset sizes varying
454 from 100 to 20,000. These features were included in training as CpG² beta values, and, when not
455 already present in the model, in their linear form as well. In addition to the top 10,000 age-associated
456 CpGs, the top 300 quadratic sites from our EWAS yielded the best performing model (**Supplementary**
457 **Table 4, Supplementary Figure 5**). This final list of features was then trained and tested using a LOCO
458 framework, as described above.

459
460 While this involves substantial overfitting in the training data, the test sets (other than GSE40279)
461 remained completely independent prior to the prediction analyses.

462
463 EWAS of all-cause mortality
464
465 An EWAS was conducted to identify CpG sites (from a total of 752,722 loci) that were associated with
466 time to all-cause mortality in Generation Scotland. Cox Proportional Hazards (Cox PH) regression
467 models were fit for each CpG site as predictor of interest using the *coxph* function from the *survival* R
468 package (version 3.3.1), with time-to-death or censoring as the survival outcome. Fixed effect
469 covariates included those used in our cAge EWAS (age at baseline, sex, set/batch, smoking status,
470 smoking pack years, WBC estimates, and top 20 methylation PCs). Epigenome-wide significance was
471 set at *p*-value < 3.6 x 10⁻⁸.

472
473 To assess whether relatedness in the cohort influenced the results, a Cox PH model with a kinship
474 matrix was fit for each significantly associated CpG, using the *coxme* R package (version 2.2.16). All
475 associations were replicated at *p* < 3.6 x 10⁻⁸ (**Supplementary Table 8**).

476
477 Prediction of survival (biological age)

478

479 *Training in Generation Scotland*

480

481 To train a bAge predictor, component scores for GrimAge were estimated for all Generation Scotland
482 samples via Horvath's online calculator¹⁷ (<http://dnamage.genetics.ucla.edu/new>). These included
483 DNAm estimates of smoking and seven proteins – DNAm ADM, DNAm B2M, DNAm cystatin C, DNAm
484 GDF15, DNAm leptin, DNAm PAI1, and DNAm TIMP1. Each variable was then standardised to have a
485 mean of zero and variance of one. We also considered DNAm EpiScores for 109 proteins as described
486 by Gadd et al¹⁴. The 109 EpiScores were projected into Generation Scotland via the MethylDetectR⁹⁰
487 Shiny App (<https://shiny.igmm.ed.ac.uk/MethylDetectR/>) before being standardised to have a mean of
488 zero and variance of one.

489

490 This resulted in 116 protein EpiScores, a smoking EpiScore, plus chronological age and sex as features
491 for an elastic net Cox PH model (R package *glmnet* version 4.1.4). 20-fold CV was performed (with
492 approximately 1,000 individuals per fold), with individuals from the same batch/set included in the same
493 fold, and with Harrell's C index used to evaluate the optimal λ value.

494

495 *Testing in LBC, FHS, and WHI*

496

497 The association between bAgeAccel (the residual of bAge regressed on chronological age to obtain
498 measure of accelerated epigenetic ageing) and mortality was assessed in six datasets from four
499 external studies: LBC1921 and LBC1936, FHS, and the WHI studies for White, Black, and Hispanic
500 ancestries (**Table 2**). After generating the bAge predictors in the external datasets, Cox proportional
501 hazards models, adjusting for age and sex, were used to compare associations with all-cause mortality
502 for GrimAgeAccel and bAgeAccel. We examined Schoenfeld residuals in the LBC models to check the
503 proportional hazards assumption at both global and variable-specific levels using the *cox.zph* function
504 from the R *survival* package (version 3.3.1). We restricted the TTE period by each year of possible
505 follow-up, from 5 to 21 years, and found minimal differences in the bAgeAccel-survival HRs between
506 follow-up periods that did not violate the assumption and those that did (**Supplementary Table 12**).

507

508 Enrichment analyses

509

510 A gene set enrichment analysis was performed using the Functional Mapping and Annotation (FUMA)
511 GENE2FUNC tool⁹¹, which employs a hypergeometric test. Background genes employed included all
512 unique genes tagged by CpGs in the EPIC array. FDR *p*-value threshold was set at 0.05, and the
513 minimum number of overlapping genes within gene sets was set to 2.

References

1. Yousefi, P. D. *et al.* DNA methylation-based predictors of health: applications and statistical considerations. *Nat. Rev. Genet.* **2022** *23*, 369–383 (2022).
2. Bocklandt, S. *et al.* Epigenetic predictor of age. *PLoS One* **6**, (2011).
3. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, 1–20 (2013).
4. Hannum, G. *et al.* Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Mol. Cell* **49**, 359–367 (2013).
5. Zhang, Q. *et al.* Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. *Genome Med.* **11**, 1–11 (2019).
6. Field, A. E. *et al.* DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. *Mol. Cell* **71**, 882 (2018).
7. Horvath, S. & Raj, K. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nat. Rev. Genet.* **2018** *196* **19**, 371–384 (2018).
8. Bell, C. G. *et al.* DNA methylation aging clocks: challenges and recommendations. *Genome Biol.* **2019** *201* **20**, 1–24 (2019).
9. McCartney, D. L. *et al.* Epigenetic signatures of starting and stopping smoking. *EBioMedicine* **37**, 214–220 (2018).
10. Joehanes, R. *et al.* Epigenetic Signatures of Cigarette Smoking. *9*, 436–447 (2016).
11. McCartney, D. L. *et al.* Epigenetic prediction of complex traits and death. *Genome Biol.* **19**, 136 (2018).
12. Liu, C. *et al.* A DNA methylation biomarker of alcohol consumption. *Mol. Psychiatry* **23**, 422–433 (2018).
13. Wahl, S. *et al.* Epigenome-wide association study of body mass index, and the adverse outcomes of adiposity. *Nature* **541**, 81–86 (2017).
14. Gadd, D. A. *et al.* Epigenetic scores for the circulating proteome as tools for disease prediction. *Elife* **11**, (2022).
15. Levine, M. E. *et al.* An epigenetic biomarker of aging for lifespan and healthspan. *Aging (Albany. NY)* **10**, 573–591 (2018).
16. Belsky, D. W. *et al.* Quantification of the pace of biological aging in humans through a blood test, the DunedinPoAm DNA methylation algorithm. *Elife* **9**, 1–56 (2020).
17. Lu, A. T. *et al.* DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (Albany. NY)* **11**, 303–327 (2019).
18. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Statistical Methodol.)* **67**, 301–320 (2005).
19. Smith, B. H. *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.* **42**, 689–700 (2013).
20. Bojesen, S. E., Timpson, N., Relton, C., Davey Smith, G. & Nordestgaard, B. G. AHRR (cg05575921) hypomethylation marks smoking behaviour, morbidity and mortality. *Thorax* **72**, 646–653 (2017).
21. Zhang, Y., Yang, R., Burwinkel, B., Breitling, L. P. & Brenner, H. F2RL3 methylation as a biomarker of current and lifetime smoking exposures. *Environ. Health Perspect.* **122**, 131–137 (2014).
22. Haase, T. *et al.* Novel DNA Methylation Sites Influence GPR15 Expression in Relation to Smoking. *8*, 74 (2018).
23. Guida, F. *et al.* Dynamics of smoking-induced genome-wide methylation changes with time since smoking cessation. *Hum. Mol. Genet.* **24**, 2349–2359 (2015).
24. Sun, Y. Q. *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.* **50**, 1482 (2021).
25. Mathios, D. *et al.* Genome-wide investigation of intragenic DNA methylation identifies ZMIZ1 gene as a

prognostic marker in glioblastoma and multiple cancer types. *Int. J. cancer* **145**, 3425–3435 (2019).

26. Feng, Y. *et al.* SOCS3 Promoter Hypermethylation Is a Favorable Prognosticator and a Novel Indicator for G-CIMP-Positive GBM Patients. *PLoS One* **9**, e91829 (2014).

27. Liu, K., Wu, Z., Chu, J., Yang, L. & Wang, N. Promoter methylation and expression of SOCS3 affect the clinical outcome of pediatric acute lymphoblastic leukemia by JAK/STAT pathway. *Biomed. Pharmacother.* **115**, 108913 (2019).

28. Huang, L. *et al.* Transcriptional repression of SOCS3 mediated by IL-6/STAT3 signaling via DNMT1 promotes pancreatic cancer growth and metastasis. *J. Exp. Clin. Cancer Res.* **35**, 1–15 (2016).

29. Chen, Y., Tsai, Y. H. & Tseng, S. H. Regulation of ZMYND8 to Treat Cancer. *Molecules* **26**, (2021).

30. Kolla, V., Zhuang, T., Higashi, M., Naraparaju, K. & Brodeur, G. M. Role of CHD5 in human cancers: 10 years later. *Cancer Res.* **74**, 652–658 (2014).

31. Fatemi, M. *et al.* Epigenetic silencing of CHD5, a novel tumor suppressor gene, occurs in early colorectal cancer stages. *Cancer* **120**, 172 (2014).

32. Zhao, R., Meng, F., Wang, N., Ma, W. & Yan, Q. Silencing of CHD5 Gene by Promoter Methylation in Leukemia. *PLoS One* **9**, 85172 (2014).

33. 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. <https://doi.org/10.1080/15592294.2021.1980688> (2021). doi:10.1080/15592294.2021.1980688

34. Porter, L. F. *et al.* Whole-genome methylation profiling of the retinal pigment epithelium of individuals with age-related macular degeneration reveals differential methylation of the SKI, GTF2H4, and TNXB genes. *Clin. Epigenetics* **11**, 1–14 (2019).

35. Alisch, R. S. *et al.* Age-associated DNA methylation in pediatric populations. doi:10.1101/gr.125187.111

36. Johnson, N. D. *et al.* Non-linear patterns in age-related DNA methylation may reflect CD4+ T cell differentiation. *Epigenetics* **12**, 492 (2017).

37. Vershinina, O., Bacalini, M. G., Zaikin, A., Franceschi, C. & Ivanchenko, M. Disentangling age-dependent DNA methylation: deterministic, stochastic, and nonlinear. *Sci. Reports* **2021** **11** **11**, 1–12 (2021).

38. Koch, C. M. & Wagner, W. Epigenetic-aging-signature to determine age in different tissues. *Aging (Albany, NY)* **3**, 1018–1027 (2011).

39. Karir, P., Goel, N. & Garg, V. K. Human age prediction using DNA methylation and regression methods. *Int. J. Inf. Technol.* **12**, 373–381 (2020).

40. Bekaert, B., Kamalandua, A., Zapico, S. C., Van De Voorde, W. & Decorte, R. Improved age determination of blood and teeth samples using a selected set of DNA methylation markers. *Epigenetics* **10**, 922–930 (2015).

41. Choi, H., Joe, S. & Nam, H. Development of Tissue-Specific Age Predictors Using DNA Methylation Data. *Genes (Basel)* **10**, (2019).

42. Xu, C. *et al.* A novel strategy for forensic age prediction by DNA methylation and support vector regression model. *Sci. Reports* **2015** **5** **1**, 1–10 (2015).

43. Boroni, M. *et al.* Highly accurate skin-specific methylome analysis algorithm as a platform to screen and validate therapeutics for healthy aging. *Clin. Epigenetics* **12**, 1–16 (2020).

44. Everson, T. M. *et al.* DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. *Genome Med.* **7**, (2015).

45. Baur, B. & Bozdag, S. A Feature Selection Algorithm to Compute Gene Centric Methylation from Probe Level Methylation Data. *PLoS One* **11**, e0148977 (2016).

46. 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). doi:10.1101/2022.04.02.486242

47. Higgins-Chen, A. T. *et al.* A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking. *Nat. Aging* **2022** 1–18 (2022). doi:10.1038/s43587-022-00248-2

48. Sugden, K. *et al.* Patterns of Reliability: Assessing the Reproducibility and Integrity of DNA Methylation Measurement. *Patterns* **1**, 100014 (2020).
49. Logue, M. W. *et al.* The correlation of methylation levels measured using Illumina 450K and EPIC BeadChips in blood samples. *Epigenomics* **9**, 1363–1371 (2017).
50. Taylor, J. & Tibshirani, R. Post-selection inference for -penalized likelihood models. *Can. J. Stat.* **46**, 41–61 (2018).
51. de Lima Camillo, L. P., Lapierre, L. R. & Singh, R. A pan-tissue DNA-methylation epigenetic clock based on deep learning. *npj Aging* **2022** *81* **8**, 1–15 (2022).
52. Battram, T. *et al.* The EWAS Catalog: a database of epigenome-wide association studies. *Wellcome Open Res.* **2022** *741* **7**, 41 (2022).
53. McCartney, D. L. *et al.* An epigenome-wide association study of sex-specific chronological ageing. *Genome Med.* **12**, 1–11 (2019).
54. Thymus Regeneration, Immunorestoration, and Insulin Mitigation Extension Trial - Full Text View - ClinicalTrials.gov. Available at: <https://clinicaltrials.gov/ct2/show/NCT04375657>. (Accessed: 4th July 2022)
55. Lu, Y. *et al.* Reprogramming to recover youthful epigenetic information and restore vision. *Nat.* **2020** *5887836* **588**, 124–129 (2020).
56. Olova, N., Simpson, D. J., Marioni, R. E. & Chandra, T. Partial reprogramming induces a steady decline in epigenetic age before loss of somatic identity. *Aging Cell* **18**, (2019).
57. Strandberg, T. E. & Tilvis, R. S. C-Reactive Protein, Cardiovascular Risk Factors, and Mortality in a Prospective Study in the Elderly. *Arterioscler. Thromb. Vasc. Biol.* **20**, 1057–1060 (2000).
58. Lobo, S. M. A. *et al.* C-reactive protein levels correlate with mortality and organ failure in critically ill patients. *Chest* **123**, 2043–2049 (2003).
59. Mendall, M. A. *et al.* C-reactive protein: relation to total mortality, cardiovascular mortality and cardiovascular risk factors in men. *Eur. Heart J.* **21**, 1584–1590 (2000).
60. Furman, D. *et al.* Chronic inflammation in the etiology of disease across the life span. *Nat. Med.* **2019** *2512* **25**, 1822–1832 (2019).
61. Chung, H. Y. *et al.* Redefining Chronic Inflammation in Aging and Age-Related Diseases: Proposal of the Senoinflammation Concept. *Aging Dis.* **10**, 367 (2019).
62. 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.* **72**, 1054–1061 (2017).
63. Junnila, R. K. *et al.* Disruption of the GH Receptor Gene in Adult Mice Increases Maximal Lifespan in Females. *Endocrinology* **157**, 4502–4513 (2016).
64. Bartke, A. Growth Hormone and Aging: Updated Review. *World J. Mens. Health* **37**, 19 (2019).
65. Aguiar-Oliveira, M. H. & Bartke, A. Growth Hormone Deficiency: Health and Longevity. *Endocr. Rev.* **40**, 575–601 (2019).
66. Laron, Z. Do deficiencies in growth hormone and insulin-like growth factor-1 (IGF-1) shorten or prolong longevity? *Mech. Ageing Dev.* **126**, 305–307 (2005).
67. Gao, X., Jia, M., Zhang, Y., Breitling, L. P. & 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* **7**, 1–10 (2015).
68. Philibert, R. A., Dogan, M. V., Mills, J. A. & Long, J. D. AHRR Methylation is a Significant Predictor of Mortality Risk in Framingham Heart Study. *J. Insur. Med.* **48**, 79–89 (2019).
69. Zhang, Y. *et al.* F2RL3 methylation in blood DNA is a strong predictor of mortality. *Int. J. Epidemiol.* **43**, 1215–1225 (2014).
70. Zhang, Y. *et al.* DNA methylation signatures in peripheral blood strongly predict all-cause mortality. *Nat. Commun.* **2017** *81* **8**, 1–11 (2017).
71. Ono, M. *et al.* Foxp3 controls regulatory T-cell function by interacting with AML1/Runx1. *Nature* **446**, 685–689 (2007).

72. Fu, L. *et al.* High expression of ETS2 predicts poor prognosis in acute myeloid leukemia and may guide treatment decisions. *J. Transl. Med.* **15**, 1–9 (2017).
73. Liquori, A. *et al.* Acute Promyelocytic Leukemia: A Constellation of Molecular Events around a Single PML-RARA Fusion Gene. *Cancers (Basel.)* **12**, (2020).
74. McCartney, D. L. *et al.* Investigating the relationship between DNA methylation age acceleration and risk factors for Alzheimer's disease. *Alzheimer's Dement. Diagnosis, Assess. Dis. Monit.* **10**, 429–437 (2018).
75. McCartney, D. L. *et al.* Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip. *Genomics Data* **9**, 22 (2016).
76. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 1–10 (2013).
77. Deary, I. J., Gow, A. J., Pattie, A. & Starr, J. M. Cohort profile: the Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **41**, 1576–1584 (2012).
78. Taylor, A. M., Pattie, A. & Deary, I. J. Cohort Profile Update: The Lothian Birth Cohorts of 1921 and 1936. *Int. J. Epidemiol.* **47**, 1042–1060 (2018).
79. Horvath, S. *et al.* An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol.* **17**, 1–23 (2016).
80. Horvath, S. *et al.* Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol.* **13**, R97 (2012).
81. 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 (2013).
82. 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.* **10**, e1004211 (2014).
83. Marioni, R. E. *et al.* DNA methylation age of blood predicts all-cause mortality in later life. *Genome Biol.* **16**, 1–12 (2015).
84. DAWBER, T. R., MEADORS, G. F. & MOORE, F. E. Epidemiological Approaches to Heart Disease: The Framingham Study. *Am. J. Public Heal. Nations Heal.* **41**, 279 (1951).
85. Kannel, W. B., Feinleib, M., McNamara, P. M., Garrison, R. J. & Castelli, W. P. An investigation of coronary heart disease in families. The Framingham offspring study. *Am. J. Epidemiol.* **110**, 281–290 (1979).
86. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363 (2014).
87. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 1–16 (2012).
88. Saffari, A. *et al.* Estimation of a significance threshold for epigenome-wide association studies. *Genet. Epidemiol.* **42**, 20–33 (2018).
89. Zeng, Y. & Breheny, P. The biglasso Package: A Memory- and Computation-Efficient Solver for Lasso Model Fitting with Big Data in R. *R J.* **12**, 1–14 (2017).
90. Hillary, R. F. & Marioni, R. E. MethylDetectR: A software for methylation-based health profiling. *Wellcome Open Res.* **5**, (2021).
91. Watanabe, K., Taskesen, E., Van Bochoven, A. & Posthuma, D. Functional mapping and annotation of genetic associations with FUMA. *Nat. Commun.* **2017** *81* **8**, 1–11 (2017).

Ethics

All components of Generation Scotland received ethical approval from the NHS Tayside Committee on Medical Research Ethics (REC Reference Number: 05/S1401/89). Generation Scotland has also been granted Research Tissue Bank status by the East of Scotland Research Ethics Service (REC Reference Number: 20-ES-0021), providing generic ethical approval for a wide range of uses within medical research.

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.

Availability of data and material

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.

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: https://github.com/elenabernabeu/cage_bage

EWAS summary statistics will be submitted to the EWAS catalog upon acceptance. They are currently available for open access on Edinburgh DataShare: <https://datashare.ed.ac.uk/handle/10283/4496>

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

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

Visualization of CpG-age relationships can be viewed using MethylBrowsR:

<https://shiny.igmm.ed.ac.uk/MethylBrowsR/>

Competing interests

R.E.M has received a speaker fee from Illumina and is an advisor to the Epigenetic Clock Development Foundation and Optima Partners. R.F.H. has received consultant fees from Illumina. 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.

Funding

Generation Scotland: Generation Scotland received core support from the Chief Scientist Office of the Scottish Government Health Directorates (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 [108890/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 PCEGP3-181181 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 (104036/Z/14/Z, 108890/Z/15/Z, and 221890/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.

Table 1. Age profile and test set prediction accuracy of cohorts used in cAge predictor training and testing. External cohort information taken from Zhang et al⁵. 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 _{Females} (%)	Tissue	Prediction Accuracy		
						r	RMSE	MAE
GS	18,413	47.5 (14.9)	[17.1, 98.5]	10,833 (58.8%)	Blood	-	-	-
LBC1921 ^{77,78}	692	82.3 (4.3)	[77.8, 90.6]	401 (57.9%)	Blood	0.659	4.050	2.466
LBC1936	2,795	73.6 (3.7)	[67.7, 80.9]	1,356 (48.5%)	Blood	0.685	3.311	2.099
GSE72775 ⁷⁹	335	70.2 (10.3)	[36.5, 90.5]	138 (41.2%)	Blood	0.949	3.275	1.843
GSE78874 ⁷⁹	259	68.8 (9.7)	[36.0, 88.0]	113 (43.6%)	Saliva	0.875	6.826	4.333
GSE72773 ⁷⁹	310	65.6 (13.9)	[35.1, 91.9]	150 (48.4%)	Blood	0.945	4.611	2.068
GSE72777 ⁷⁹	46	14.7 (10.4)	[2.2, 35.0]	31 (67.4%)	Blood	0.942	4.211	2.505
GSE41169 ^{a,80}	95	31.6 (10.3)	[18.0, 65.0]	28 (29.5%)	Blood	0.975	2.869	1.947
GSE40279 ⁴	656	64.0 (14.7)	[19.0, 101.0]	338 (51.5%)	Blood	0.969	3.697	2.074
GSE42861 ^{a,81}	689	51.9 (11.8)	[18.0, 70.0]	492 (71.4%)	Blood	0.972	4.498	3.563
GSE53740 ^{a,82}	383	67.8 (9.6)	[34.0, 93.0]	155 (40.5%)	Blood	0.921	4.443	2.797

^a 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 **Supplementary Table 11**. *The FHS cohort used here was the same as the test set from the original GrimAge paper.

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

Figure 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 visualized 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/elenabernabeu/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/elenabernabeu/cage_bage/tree/main/bage_predictor), after obtaining GrimAge estimates from an external online calculator (<http://dnamage.genetics.ucla.edu/new>).

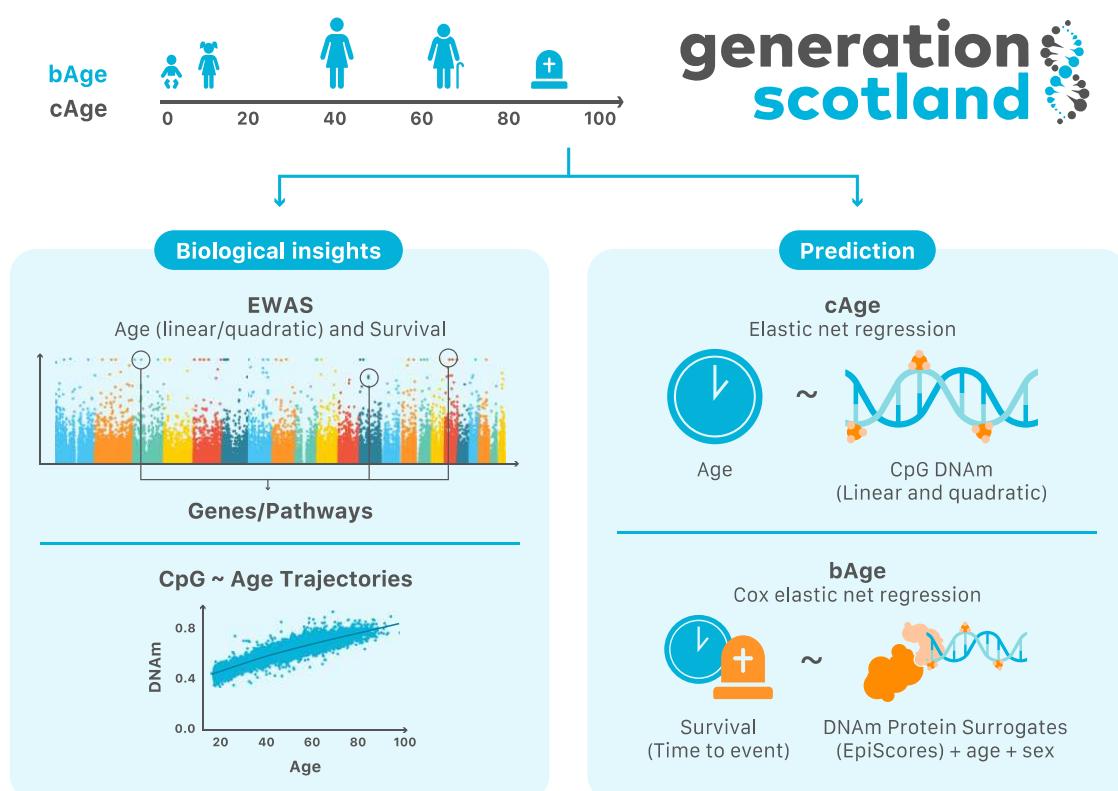


Figure 2. Flowchart for the creation of the cAge predictor. First, DNAm data originating from Generation Scotland and 10 external datasets was pre-processed. Next, CpGs were pre-selected based on the Generation Scotland EWAS for genome-wide significant linear and quadratic CpG-age associations. Elastic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

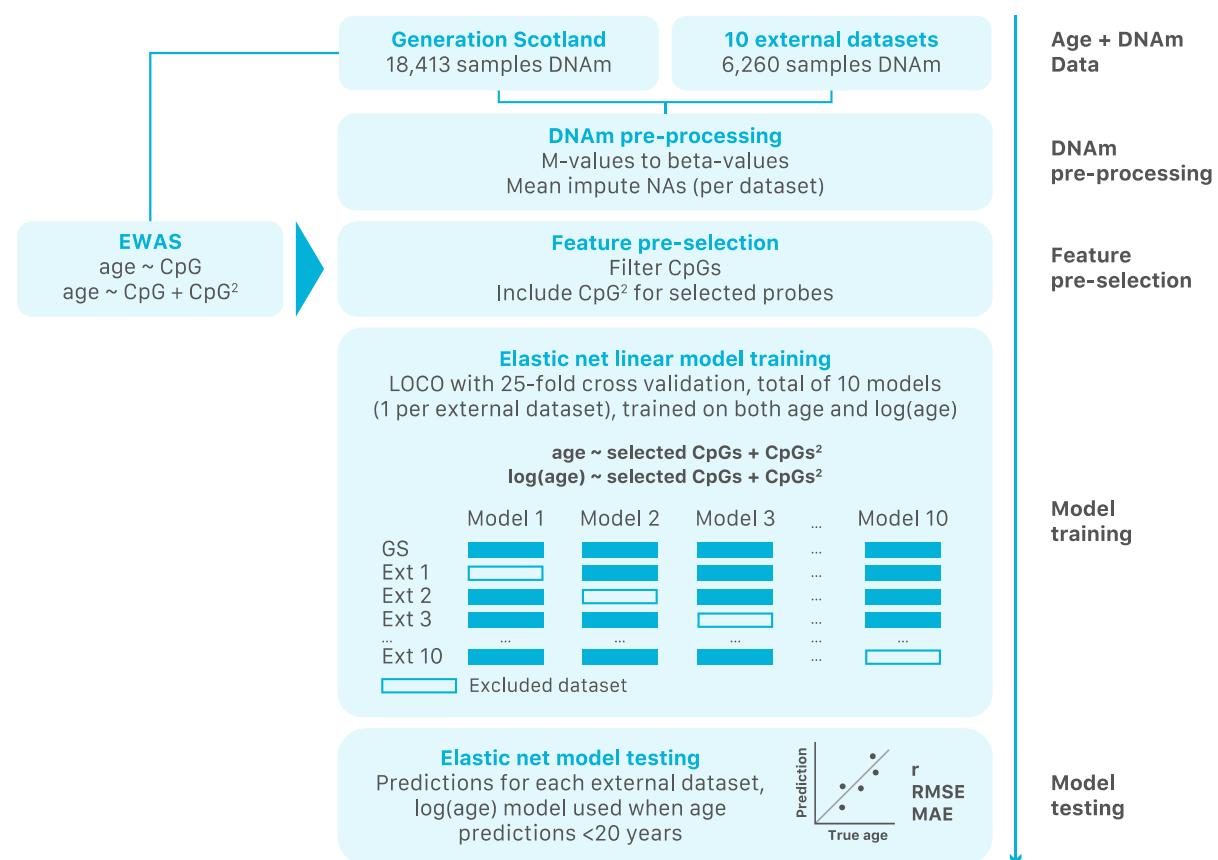


Figure 3. cAge predictor performance on 10 external testing datasets, (a) across all datasets considered, and (b) per cohort. Performance metrics shown include Pearson correlation (r), root mean squared error (RMSE), and median absolute error (MAE). Metrics also included in **Table 1**.

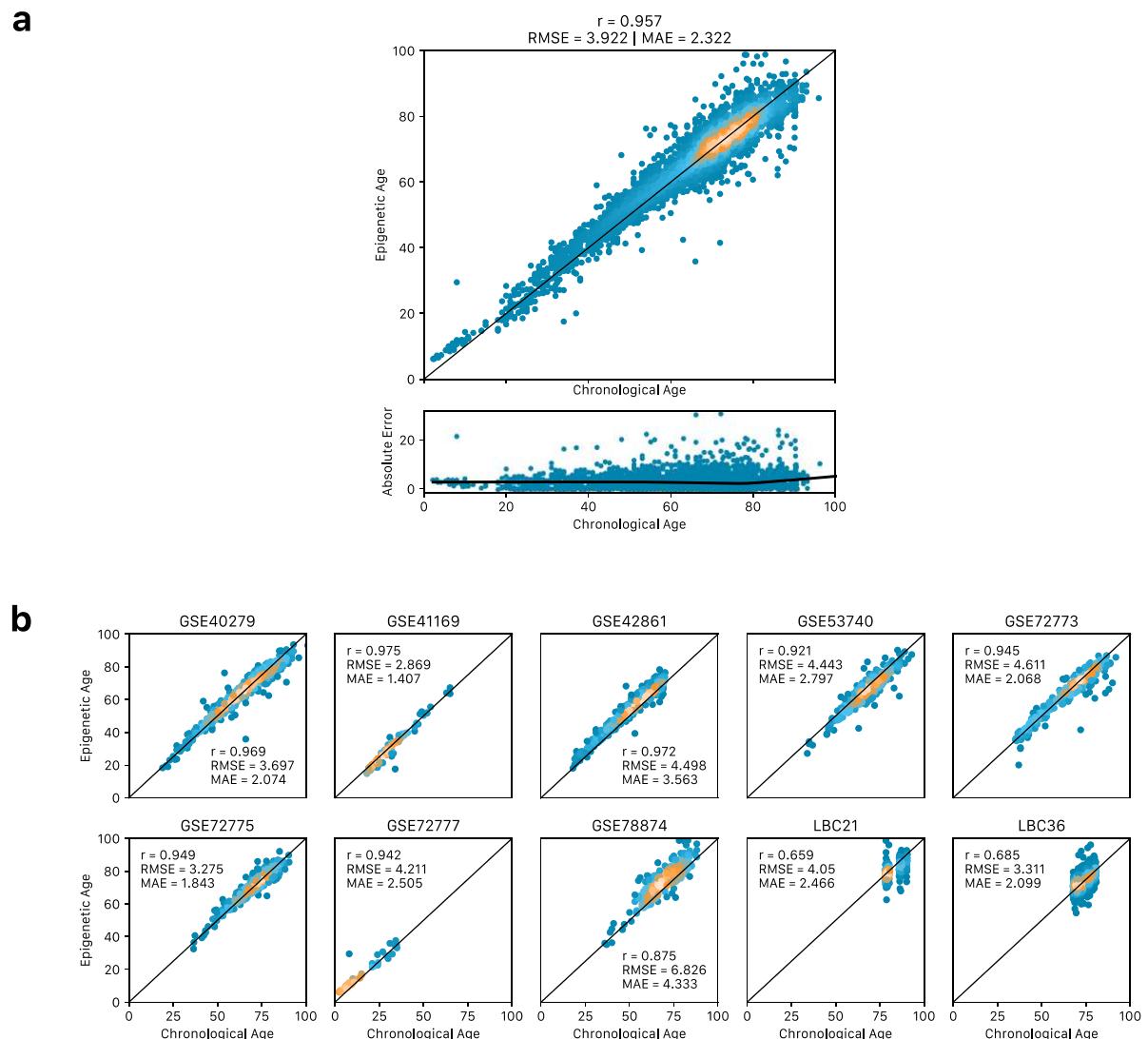


Figure 4. Flowchart for the creation of the bAge predictor. First, DNAm data originating from Generation Scotland and six external datasets was pre-processed. GrimAge components and 109 protein EpiScores were generated within each cohort. A Cox proportional hazards elastic net regression model of all-cause mortality (with 20-fold cross validation) 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.

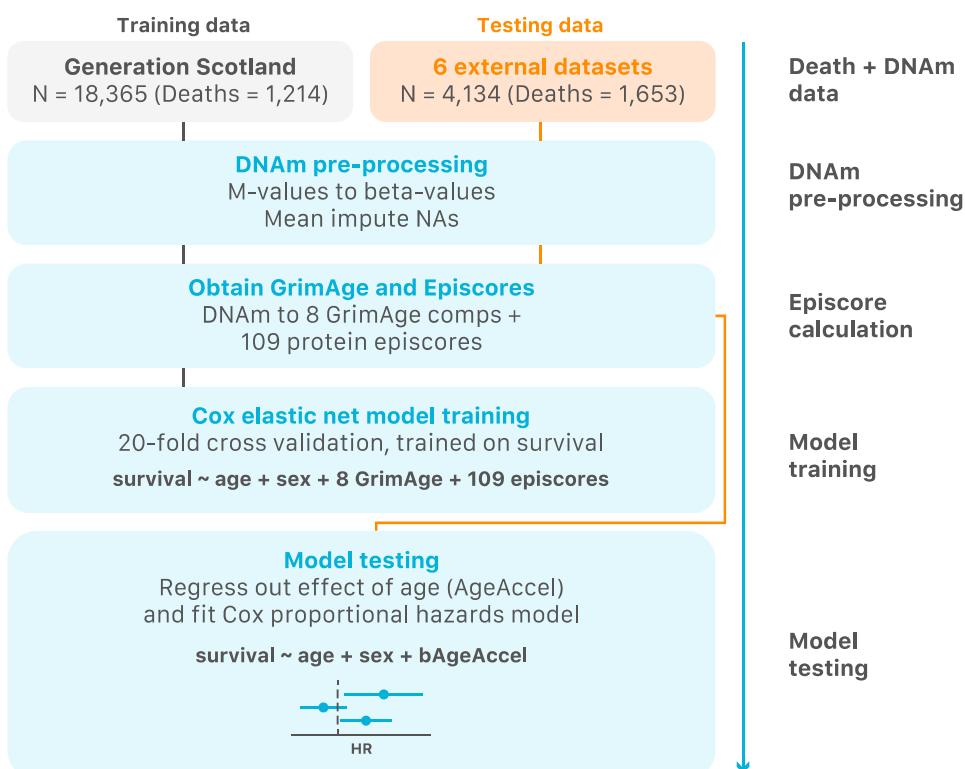


Figure 5. Forest plots of bAge/GrimAge predictors, applied to all-cause mortality in LBC1921, LBC1936, FHS, and WHI. Predictors regressed on age. Hazard ratios are presented per standard deviation of the GrimAgeAccel and bAgeAccel variables, along with 95% confidence intervals. Cox models are adjusted for age at DNAm sampling and sex.

