nature aging

Technical Report

https://doi.org/10.1038/s43587-026-01073-7

OMICmAge quantifies biological age by integrating multi-omics with electronic medical records

Received: 27 November 2023

Accepted: 12 January 2026

Published online: 25 February 2026

Check for updates

Qingwen Chen[1,2,20], Varun B. Dwaraka[3,20], Natàlia Carreras-Gallo[3], Jenel F. Armstrong[4], Raghav Sehgal[5], M. Austin Argentieri[6,7], Anne Richmond[8], Andrea Aparicio[1], Kevin Mendez[1], Yulu Chen[1], Sofina Begum[1], Priyadarshini Kachroo[1,9], Nicole Prince[1], Tao Guo[1], Hannah Went[3], Tavis Mendez[3], Aaron Lin[3], Logan Turner[3], Mahdi Moqri[10,11,12], Su H. Chu[1], Rachel S. Kelly[1], Scott T. Weiss[1], Nicholas J. W. Rattray[13,14], Vadim N. Gladyshev[10], Elizabeth Karlson[15], Craig E. Wheelock[16,17], Ewy A. Mathé[18], Amber Dahlin[1], Michael J. McGeachie[1], Riccardo E. Marioni[8], Albert T. Higgins-Chen[4,5,19], Ryan Smith[3] & Jessica Lasky-Su[1]

Biological aging reflects complex cellular and biochemical processes that can be measured across multiple omic layers. Using routine clinical laboratory data from ~31,000 participants in the Mass General Brigham Biobank, we developed EMRAge, a biomarker of mortality risk that can be broadly recapitulated across electronic medical records. Here we show that EMRAge can be modeled using elastic net regression with DNA methylation and multi-omics to generate DNAmEMRAge and OMICmAge, respectively. Both biomarkers are strongly associated with incident and prevalent chronic diseases and mortality, performing comparably or better than current biomarkers across discovery (Massachusetts General Brigham Aging Biobank Cohort, $n = 3,451$) and validation cohorts (TruDiagnostic, $n = 14,213$; Generation Scotland, $n = 18,672$). Importantly, OMICmAge leverages epigenetic biomarker proxies to integrate proteomic, metabolomic and clinical domains while remaining quantifiable from DNA methylation alone. This framework establishes an accessible, scalable measure of biological aging with potential to reveal molecular interconnections that shape healthspan and disease risk.

A major goal of aging research is to define biomarkers of aging that capture interindividual differences in functional decline, chronic disease development and mortality not identified through chronological age alone[1]. Molecular and clinical data quantify complementary attributes of biological aging. Multiple molecular biomarkers of aging, or 'clocks', have been developed as proxies for these hallmarks of aging[2]. These biomarkers have been based on a variety of measures such as telomere length[3], neuroimaging data[4–7], immune cell counts[8] and large-scale omics including DNA methylation (DNAm)[2,9–11], metabolomics[12], glycomics[13] and proteomics[14–17].

Over the last two decades, electronic medical records (EMRs) have been widely used in clinical research, in particular for precision medicine, enabling deep phenotype mining from dense, comprehensive time-dependent data[18]. These data track longitudinal physiological change and real-time health status. Capitalizing on EMRs provides a unique opportunity to quantify the aging process in a reproducible way

A full list of affiliations appears at the end of the paper. e-mail: rejas@channing.harvard.edu

across clinical settings. While healthy aging encompasses both quality of life and lifespan, metrics of biological age have traditionally focused on using either clinical data to quantify quality of life[19,20], or mortality risk to quantify lifespan[21], resulting in biological phenotypes that are optimized to one of these attributes, while not fully reflecting the other. With the wealth of data available via EMRs, biological aging phenotypes that incorporate both dense clinical data and mortality can be created to synthesize these important attributes of aging into a single measure.

Clinical data are essential in constructing and validating age readouts; connecting them to molecular underpinnings is equally important. Here we combine EMR data with multi-omic profiling to develop a more biologically informed measure of aging. The strong molecular link between DNAm and aging has driven widespread development of DNAm clocks reflecting clinical biomarkers (for example, PhenoAge[19]), mortality (for example, GrimAge[21]) and the rate of aging (for example, DunedinPACE[22]).

Proteomics and metabolomics directly reflect biological processes that inform the aging process. The proteome is altered by hallmarks of aging including loss of proteostasis, dysregulated nutrient sensing, altered intercellular communication and cellular senescence[23]. Although the source of circulating proteins and metabolites is often unclear, individual blood-based proteins and metabolites are established biomarkers for specific organ function (for example, albumin, C-reactive protein and creatinine), while peripheral omic signatures reflect organ-specific changes with aging[17,23,24]. The metabolome not only provides critical information about metabolic processes, but also captures measures of environmental exposures, including xenobiotics, that may be critically linked to the aging process[25,26]. Blood metabolomics captures molecules from multiple tissues across the body, providing rich information on aging information that may not be captured in methylation and transcriptomic clocks[27,28].

Despite the important advantages of other omics, the development of epigenomic, proteomic and metabolomic clocks for biologic aging phenotypes has been limited. Initial work has demonstrated that while individual omics clocks share commonalities, each omic data type provides a distinct window on the aging process[29], suggesting that the best and most clinically informative approach would integrate information from multiple omic measurements to create an optimized aging biomarker. However, the integration of multiple omics into a multi-omic clock or to inform DNAm-based readouts remains an area of unfulfilled clinical potential.

To this end, we used ~31,000 participants from the Massachusetts General Brigham (MGB) Biobank to develop and validate three distinct and clinically relevant measures of biological age: (1) EMRAge, a clinically based mortality predictor that can be broadly recapitulated across EMRs; (2) DNAmEMRAge, a DNAm aging biomarker trained to predict EMRAge; and (3) OMICmAge, a multi-omic-informed aging biomarker trained to predict EMRAge, using proteomic, metabolomic and clinical data distilled into DNAm via epigenetic biomarker proxies (EBPs). Overall, these aging biomarkers show strong associations with both incident and prevalent chronic disease outcomes and mortality, while further substantiating the biological relevance and value of integrating multiple omics data into one biological aging biomarker. To validate the findings, we used three independent cohorts—All of Us, TruDiagnostic Biobank and Generation Scotland.

Results

Overview of study design

We developed and validated three aging biomarkers: EMRAge, DNAmEMRAge and OMICmAge. Participants in the MGB Biobank with available plasma and clinical data were used to develop EMRAge (n = 31,264; Extended Data Fig. 1), which was validated using the All of Us cohort (n = 10,769). A subset of individuals from the MGB cohort also had available multi-omic data and were used to develop DNAmEMRAge and OMICmAge (Massachusetts General Brigham Aging Biobank Cohort or MGB-ABC; n = 3,451). These aging biomarkers were validated using two independent cohorts—the TruDiagnostic Biobank (n = 14,213) and the Generation Scotland (n = 18,672; Fig. 1). Additional clinical characteristics and demographics are in Supplementary Table 1.

Development of EMRAge

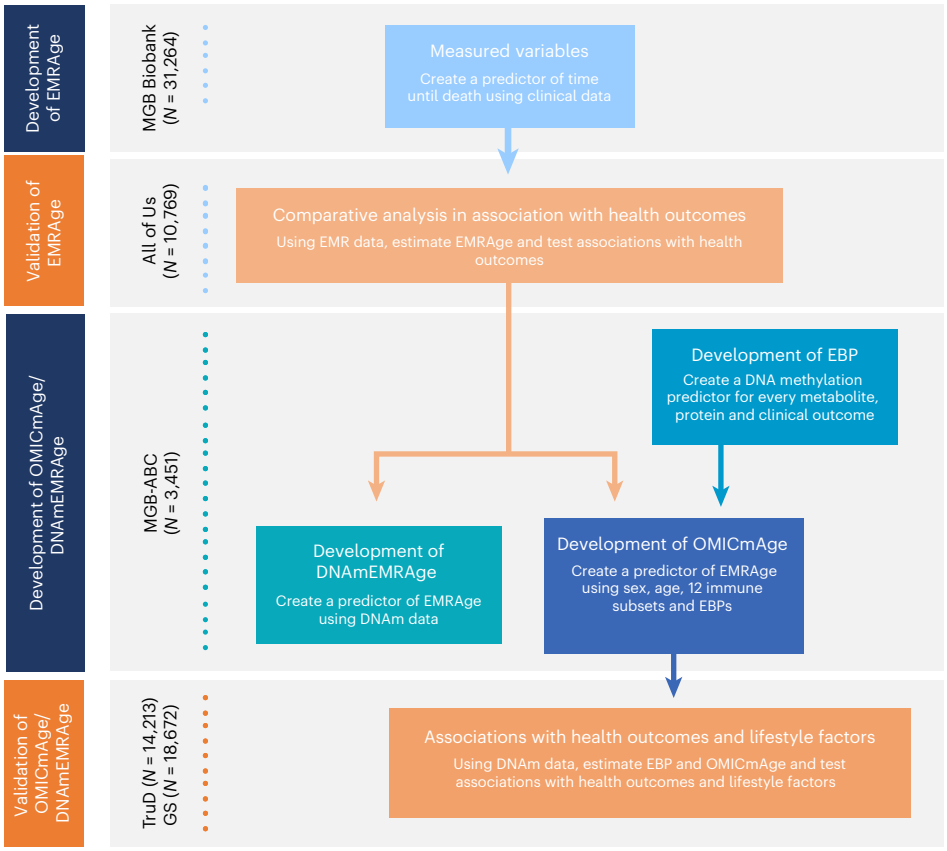
We filtered 60,370 individuals and 28 clinical phenotypes (43 clinical variables; Supplementary Table 2) from the MGB Biobank down to 31,264 individuals with complete data on 19 clinical variables that were used to develop EMRAge (Extended Data Fig. 1). We split the cohort by 70:30 into training and testing sets. A Cox proportional-hazards model was fitted in the training set to estimate the weightings of the 19 selected clinical variables (Supplementary Table 3). In a manner analogous to the GrimAge approach[21], we converted the linear combination of estimated weights and predictor values into an 'age' metric. The Pearson correlation coefficient (ρ) between EMRAge and chronological age was 0.76 ($P < 0.001$) in the testing set and 0.75 ($P < 0.001$) in the training set (Extended Data Fig. 2). We validated the EMRAge predictors by retraining the algorithm at four time points in 2-year increments: 1 January 2008, 2010, 2012 and 2014. The four derived equations were then applied to participants (N = 11,673) on 1 January 2016. The Pearson correlations among these estimates were ~1, affirming robustness (Fig. 2a). We then assessed the association between EMRAge, PhenoAge and chronological age with aging-related health outcomes, including all-cause mortality, stroke, type 2 diabetes, chronic obstructive pulmonary disease (COPD), depression, cardiovascular disease (CVD) and any type of cancer. All association tests were adjusted for age (if using EMRAge or PhenoAge), sex, race, smoking status and alcohol consumption. The prospective association analysis in the testing set shows that EMRAge has the largest hazard ratios (HRs) for all-cause mortality (HR = 4.53, $P = 4.42 \times 10^{-129}$), stroke (HR = 2.00, $P = 1.20^{-22}$), COPD (HR = 2.21, $P = 4.01^{-15}$) and cancer (HR = 2.22, $P = 2.00 \times 10^{-27}$) and demonstrates comparable HRs for type 2 diabetes (HR = 2.05, $P = 2.07 \times 10^{-19}$), depression (HR = 1.59, $P = 3.89 \times 10^{-12}$) and CVD (HR = 1.99, $P = 7.27 \times 10^{-32}$) when compared to PhenoAge (HR = 2.26, $P = 1.03 \times 10^{-18}$; HR = 1.67, $P = 3.33 \times 10^{-10}$; HR = 2.00 $P = 1.00 \times 10^{-20}$, respectively; Fig. 2b and Supplementary Table 4). All associations were significant with a false discovery rate (FDR) threshold of 0.05 after adjusting for multiple testing. Kaplan–Meier plots show that EMRAge provides the best divergence of survival probabilities among 'age' groups, compared to PhenoAge and chronological age (Fig. 2c). Additionally, a comparative analysis including both EMRAge and PhenoAge into the same model revealed that EMRAge consistently shows higher HRs for the aging-related incident outcomes (Fig. 2d and Supplementary Table 5).

Validation of EMRAge

We validated EMRAge using data from the All of Us Research Program (CDR version 8). Median imputation of missing lab values occurred within ±1 year of enrollment. Following imputation, the cohort comprised 10,769 adult participants with complete data for both EMRAge and PhenoAge calculation. Among these participants, 378 (3.5%) were recorded as deceased by 1 October 2023. Demographic characteristics are detailed in Supplementary Table 1. We then performed association analyses between EMRAge, PhenoAge or chronological age with aging-related diseases. All association tests were adjusted for the same set of covariates as previously described. Our prospective analysis (Extended Data Fig. 3a and Supplementary Table 6) revealed that EMRAge demonstrated the strongest association with all-cause mortality (HR = 3.08 (per s.d., same below), $P = 4.97 \times 10^{-74}$). Similarly, our cross-sectional analysis (Extended Data Fig. 3a and Supplementary Table 7) indicated that EMRAge exhibited the strongest associations with multiple prevalent aging-related diseases, including stroke (odds ratio (OR) = 1.08, $P = 3.14 \times 10^{-62}$), COPD (OR = 1.10, $P = 3.16 \times 10^{-96}$), depression (OR = 1.07, $P = 1.48 \times 10^{-27}$) and cancer (OR = 1.10, $P = 7.32 \times 10^{-90}$), when compared to PhenoAge and chronological age. Furthermore, in a joint

https://doi.org/10.1038/s43587-026-01073-7

a



b

Study population	EMRAge		DNAmEMRAge/OMICmAge		
	MGB	All of Us	MGB-ABC	TruD	GS
N	31,264	10,769	3,451	14,213	18,672
Sex, male	13,974	4,521	1,250	8,362	7,700
Age, mean	57	58	59	53	47
Race, *N*					
White	26,260	6,952	2,802	12,114	17,996
African American	1,806	1,074	250	454	0
Asian	728	171	59	787	0
Other	1,710	964	263	858	0
Unknown	760	1,608	77	0	676

Fig. 1 | Overall study design. a, Workflow of the study. **b**, Description of the study population. TruD, TruDiagnostic. GS, Generation Scotland.

analysis including both EMRAge and PhenoAge in the same model (Extended Data Fig. 3b), EMRAge showed a stronger association with all-cause mortality in the prospective analysis (HR = 2.40, $P = 9.08 \times 10^{-22}$; Supplementary Table 8). The cross-sectional analysis from this joint model (Supplementary Table 9) further demonstrated that EMRAge exhibited stronger associations, as illustrated in Extended Data Fig. 3b.

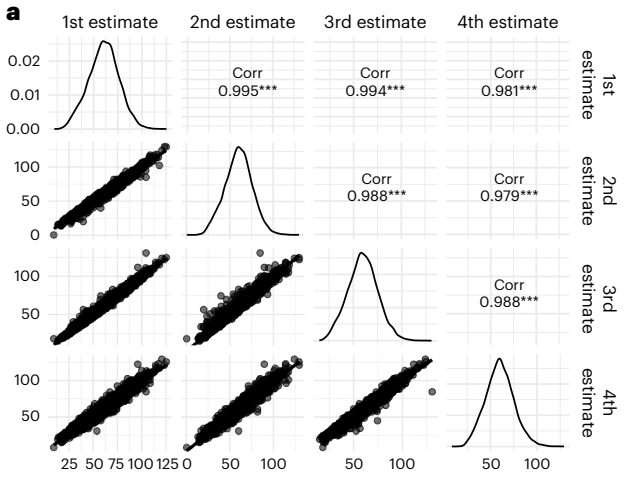
Development of DNAmEMRAge

After developing the EMRAge measure, we created a DNAm surrogate predictor of EMRAge, DNAmEMRAge, using DNAm data in an elastic net regression model (alpha = 0.1) to select the CpG sites that are most predictive of EMRAge. The model for DNAmEMRAge included 1,097 CpG sites and age as predictors. A 25-fold cross-validation selected an optimal lambda value with $R^2 = 0.827$ between the observed and predicted values, suggesting good concordance in prediction. To further assess agreement, the data were resampled to identify a training set composed of samples used to generate the model and samples not in the model ($N = 2,762$). Within the training data, DNAmEMRAge and EMRAge values showed high correlation (Fig. 3a; $N = 2,762$, $R^2 = 0.82$, $P < 2.2 \times 10^{-16}$, $\rho = 0.91$, $P < 2.2 \times 10^{-16}$). A test dataset showed comparable correlations ($N = 689$, $R^2 = 0.83$, $P < 2.2 \times 10^{-16}$, $\rho = 0.91$, $P < 2.2 \times 10^{-16}$; Fig. 3b). The mean absolute error between DNAmEMRAge and EMRAge is 8.33 years in the training set and 8.50 years in the testing set, and the intraclass correlation coefficient (ICC) was 0.995 (Fig. 3c).

Development of OMICmAge

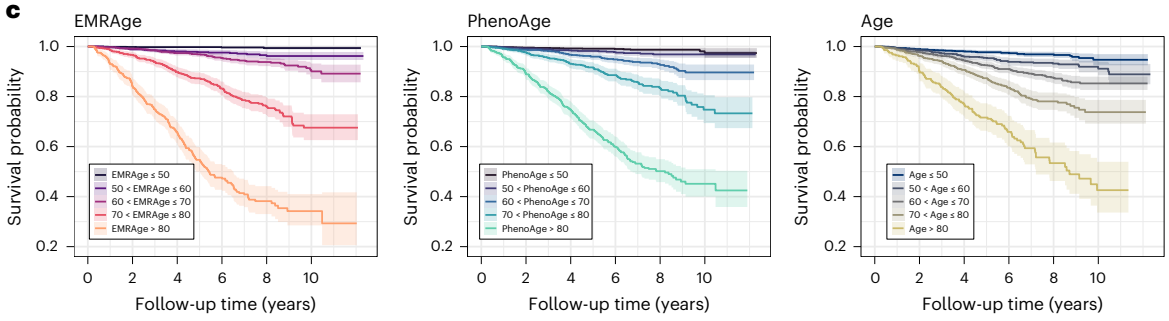
Metabolomic, proteomic and clinical EBPs. Untargeted global plasma metabolomic profiling was performed on the Metabolon platform. After preprocessing and scaling, the final metabolomic dataset consisted of 1,459 metabolites, covering a broad range of metabolic

https://doi.org/10.1038/s43587-026-01073-7



b

	N of disease-free (number of events)	HR (95% CI)		
		EMRAge	PhenoAge	Age
All-cause mortality	5,171 (496)	4.53 (4.06, 5.06)	3.95 (3.57, 4.37)	2.61 (2.31, 2.95)
Stroke	4,794 (443)	2.00 (1.76, 2.29)	1.90 (1.65, 2.17)	1.72 (1.53, 1.93)
Type 2 diabetes	4,072 (339)	2.05 (1.77, 2.37)	2.26 (1.93, 2.63)	1.22 (1.08, 1.38)
COPD	4,642 (217)	2.21 (1.84, 2.66)	2.14 (1.76, 2.60)	1.59 (1.35, 1.88)
Depression	3,549 (523)	1.59 (1.40, 1.81)	1.67 (1.44, 1.94)	0.86 (0.79, 0.94)
CVD (excluding stroke)	3,040 (699)	1.99 (1.79, 2.21)	2.00 (1.76, 2.28)	1.55 (1.42, 1.69)
Cancer	4,172 (385)	2.22 (1.94, 2.55)	1.73 (1.47, 2.03)	1.62 (1.43, 1.82)



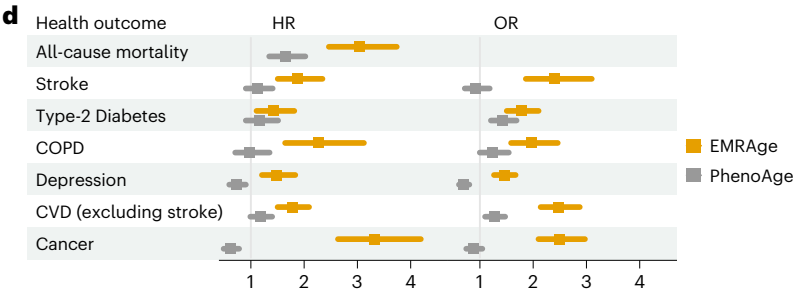


Fig. 2 | Development, robustness and comparators of EMRAge. a, Pairwise correlation (Corr) between four different estimates of EMRAge at the following time points: 1 January 2008, 2010, 2012 and 2014. ***$P < 0.001$. **b**, HRs and confidence intervals (CIs) of a 1-s.d. change to onset of aging-related diseases. These values were estimated in a subset of the testing dataset from the MGB Cohort, in which PhenoAge is available ($N = 5,171$). The models were adjusted for chronological age (if EMRAge or PhenoAge), sex, race, body mass index (BMI), smoking status and alcohol consumption. **c**, Kaplan–Meier plot of EMRAge versus PhenoAge versus chronological age in the testing dataset. Error bands show the 95% CIs for the estimated survival probabilities in each age group. **d**, Forest plot of HRs and ORs per s.d. change of EMRAge or PhenoAge for aging-related health outcomes when both variables are included as predictors in a single model, without additional adjustment for covariates. Error bars show the 95% CIs of the estimated HRs/ORs. The number of incident cases (n) and sample sizes (N) for each phenotype are as follows. Incident cases: all-cause mortality, $N = 5,171$ ($n = 496$); stroke, $N = 4,794$ ($n = 443$); type 2 diabetes, $N = 4,072$ ($n = 339$); COPD, $N = 4,642$ ($n = 217$); depression, $N = 3,549$ ($n = 523$); CVD (excluding stroke), $N = 3,040$ ($n = 699$); cancer, $N = 4,172$ ($n = 385$). Prevalent cases: The total sample size for prevalent diseases is 5,171, with the following case numbers; stroke, $n = 377$; type 2 diabetes, $n = 1,099$; COPD, $n = 529$; depression, $n = 1,622$; CVD (excluding stroke), $n = 2,131$; cancer, $n = 999$.

pathways (Extended Data Fig. 4) across 1,986 individuals, among whom 1,691 were matched to methylation data. Global proteomic data were generated using the Seer SP100 platform, based on liquid chromatography–mass spectrometry (LC–MS). The final processed dataset consisted of 2,098 nonunique annotated proteins and 536 unique protein groups (denoted as 'proteins') across 1,789 individuals, among whom 1,475 were matched with methylation data. We further considered 46 clinical variables that have potential relationships with aging and aging-related outcomes. As implemented in the development of other aging biomarkers[21], we restricted the number of EBPs for inclusion by selecting proteins and clinical variables with a significant Pearson correlation to EMRAge greater than 0.1 and a nominal P value < 0.05 (Supplementary Table 10). To select metabolites highly correlated with EMRAge while minimizing interdependence among the selected metabolites, we used hierarchical clustering. This process grouped the metabolites into 286 clusters characterized by low intercluster correlation (that is, the 90th percentile of average intercorrelations between clusters was below 0.15) and high intra-cluster correlation (that is, the 10th percentile of average intra-correlations within clusters was above 0.5). Subsequently, we selected the metabolite exhibiting the strongest correlation with EMRAge from each cluster. Following this strategy, 286 metabolites, 110 proteins and 25 clinical variables were retained (Fig. 4).

https://doi.org/10.1038/s43587-026-01073-7

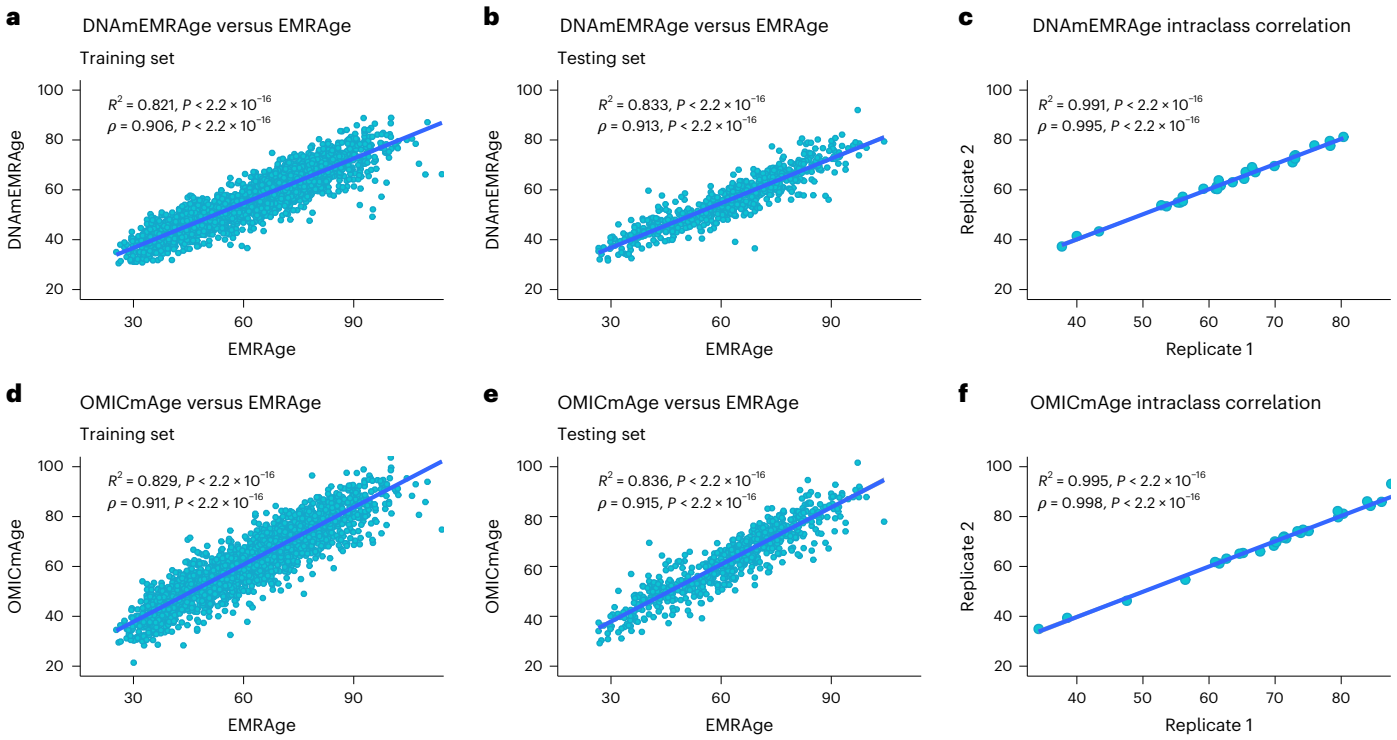


Fig. 3 | Correlation plots to EMRAge and ICCs for DNAmEMRAge and OMICmAge. The two-sided *t*-test was used to test the significance of the Pearson correlation coefficient (ρ). **a**, Correlation ($\rho = 0.906$, P value < 2.22×10^{-308}) between DNAmEMRAge and EMRAge in the training set ($N = 2{,}762$). **b**, Correlation ($\rho = 0.913$, P value = 6.26×10^{-269}) between DNAmEMRAge and EMRAge in the testing set ($N = 689$). **c**, ICCs ($\rho = 0.995$, P value = 1.45×10^{-29}) for DNAmEMRAge using 30 replicates. **d**, Correlation ($\rho = 0.911$, P value < 2.22×10^{-308}) between OMICmAge and EMRAge in the training set ($N = 2{,}762$). **e**, Correlation ($\rho = 0.915$, P value = 5.02×10^{-273}) between OMICmAge and EMRAge in the testing set ($N = 689$). **f**, ICCs ($\rho = 0.998$, P value = 3.96×10^{-35}) for OMICmAge using 30 replicates.

We then generated epigenetic predictors (that is, EBPs) for each selected metabolite, protein and clinical variable—via an elastic net regression model. We retained all EBPs with a nominal P value ($P < 0.05$) and a Pearson correlation above 0.2 with their estimated metabolite/protein/clinical value. We observed strong correlations between several of the selected EBPs and actual clinical values (for example, $\rho = 0.66$ and 0.63 for C-reactive protein and HbA1C EBPs, respectively). In total, 266 metabolite EBPs, 109 protein EBPs and 21 clinical EBPs were retained, totaling 396 EBPs that were included as features in the predictive model for OMICmAge (Supplementary Table 11). OMICmAge was then generated by integrating proteomic, metabolomic and clinical data EBPs into a DNAm clock.

Predictive model for the OMICmAge

OMICmAge was generated via a penalized elastic net regression model of EMRAge that included methylation CpG values, relative percentages of 12 immune cell subsets, 396 EBPs, age and sex as features in the model. This model retained 990 CpGs, 40 EBPs (16 protein EBPs, 14 metabolite EBPs and 10 clinical EBPs; Fig. 4) and age as selected predictors of EMRAge with varying weightings in the final model. The model did not retain any of the immune cell subsets after penalization. We tested an independent model including them as unpenalized features, but results did not change substantially. Thus, we continued with the model where all the features were penalized. Figure 3d–f shows the correlation between EMRAge and OMICmAge in the training ($N = 2{,}762$, $R^2 = 0.83$, $P < 2.2 \times 10^{-16}$; $\rho = 0.91$, $P < 2.2 \times 10^{-16}$) and testing ($N = 689$, $R^2 = 0.84$, $P < 2.2 \times 10^{-16}$; $\rho = 0.92$, $P < 2.2 \times 10^{-16}$) sets, as well as the ICCs using 30 replicates (0.998). In terms of error, the mean absolute error between OMICmAge and EMRAge was 4.96 years in the training set and 4.97 years in the testing set, lower than the mean absolute error for DNAmEMRAge (8.33 and 8.50, respectively).

Comparison of OMICmAge to previous epigenetic biomarkers of aging. We compared DNAmEMRAge and OMICmAge to previous epigenetic aging biomarkers, including DunedinPACE[22] and the principal component (PC) versions of PCHorvath[10], PCHannum[11], PCPhenoAge[19] and PCGrimAge[21] for their improved precision[30]. We compared the CpG sites included in the predictive model, and their relationship with immune cell subsets, aging-related disease outcomes and five- and ten-year mortality. Overall, we observed consistent correlations between all epigenetic clocks and immune subsets. We observed stronger correlations with sex for both OMICmAge and DNAmEMRAge ($R = 0.28$, P value = 0.02, and $R = 0.36$, P value = 0.009, respectively) when compared to previous aging biomarkers (Extended Data Fig. 5). There was minimal overlap between the CpG sites retained in DNAmEMRAge and OMICmAge and prior aging biomarkers (Fig. 5a); DNAmEMRAge and OMICmAge had 660 and 657 unique CpG sites, respectively, with 411 CpG sites shared between these measures. While PhenoAge and Horvath biomarkers share 50 CpG sites and Horvath and Hannum share 29, the maximum number of probes shared between OMICmAge and any previous aging biomarker is 3.

To estimate OMICmAge, we use the 990 CpG sites retained in the model and the 40 EBPs, which require 10,315 additional CpG sites. Remarkably, 50.8% of them (5,740) are available on the 450 K array. A version for 450 K is in development.

We compared the prevalence and incidence of aging-related disease associations between OMICmAge, DNAmEMRAge and other aging biomarkers in MGB-ABC (Fig. 5b and Supplementary Tables 12 and 13). For prevalent disease associations, OMICmAge had the highest ORs for four of six aging-related chronic diseases assessed, with particularly high ORs for type 2 diabetes (OR = 5.04, $P = 1.37 \times 10^{-15}$) and CVD (OR = 4.62, $P = 3.56 \times 10^{-12}$). The association between PCGrimAge and COPD (OR = 3.97, $P = 6.90 \times 10^{-6}$) was also particularly high. OMICmAge also had the highest ORs for stroke (OR = 2.21, $P = 1.4 \times 10^{-4}$) and

https://doi.org/10.1038/s43587-026-01073-7

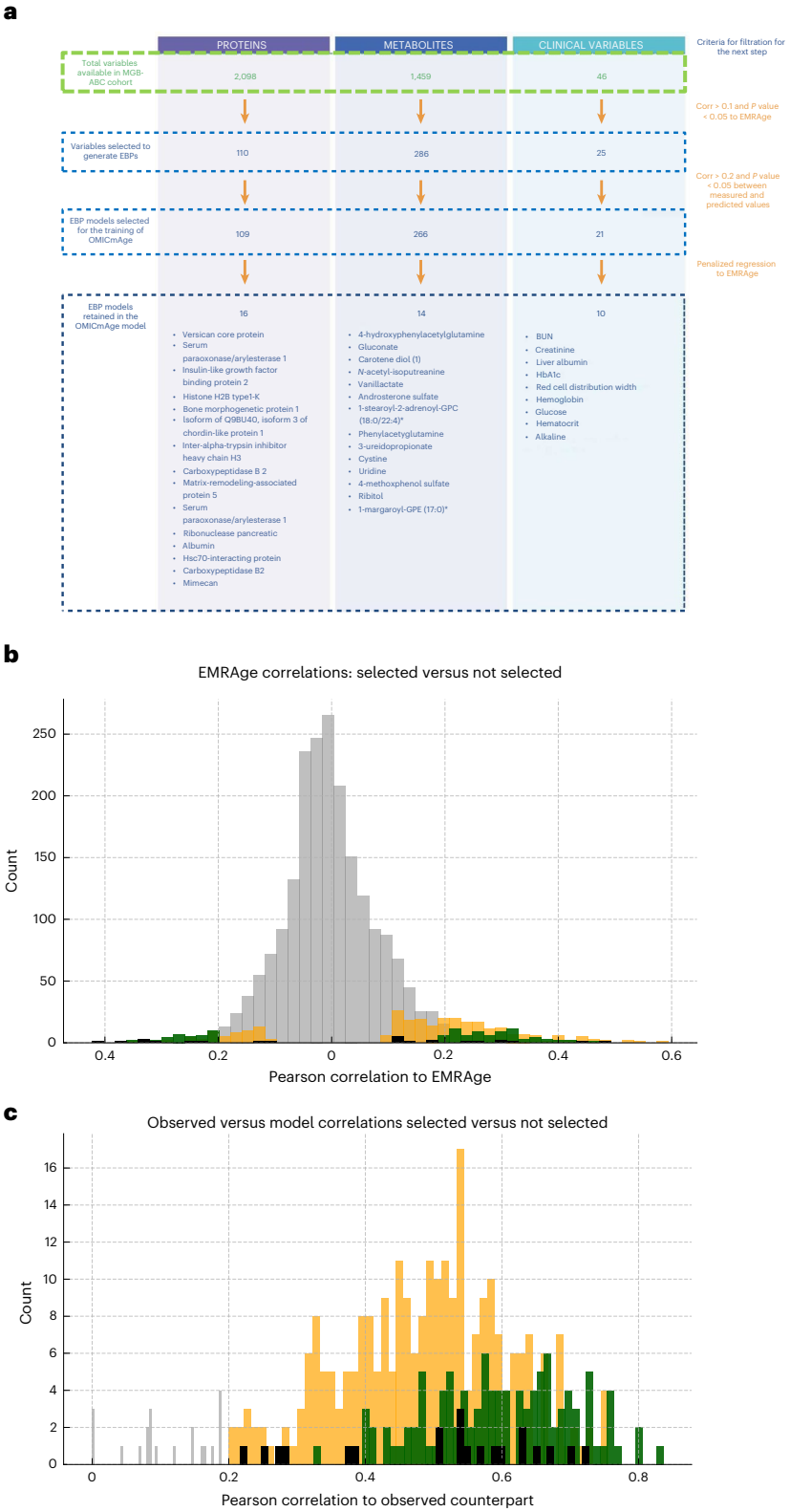


Fig. 4 | Illustration of the filtration process for DNAm-based multi-omic features used in the development of OMICmAge. a, Diagram describing the filtration and selection of EBPs included in the OMICmAge model. **b**, Histograms show the full distributions of feature-wise Pearson correlations to EMRAge for three data modalities: 2,098 proteins, 1,459 metabolites (collapsed to 286 clusters via hierarchical clustering) and 46 clinical variables. Gray bars represent the unfiltered 'background' for each modality. Colored overlays (green indicates protein EBP, orange indicates metabolite EBP, and black indicates clinical EBP) show the subset of features meeting our a priori filter ($|\rho| > 0.1$ and P value < 0.05), yielding 421 total features (protein, $n = 110$; metabolite, $n = 286$; clinical, $n = 25$). **c**, Of the 421 EMRAge-correlated candidates, we next required that predicted epigenetic biomarkers (EBPs) also correlate with their measured counterparts at Pearson $\rho > 0.2$ (P value < 0.05). Histograms again show the EMRAge-selected background distributions in gray, with colored overlays (green, protein; orange, metabolite; black, clinical) indicating the features that passed this second filter (protein, $n = 109$; metabolite, $n = 266$; clinical, $n = 21$). Pearson correlation coefficients for each multi-omic feature and EBP are reported in Supplementary Tables 10 and 11.

depression (HR = 1.94, $P = 9.77 \times 10^{-5}$) and chronological age had the highest OR for cancer (OR = 2.40, $P = 4.85 \times 10^{-13}$). However, the differences between several of the strongest aging biomarkers were all within the CIs. All these associations were significant after adjusting for multiple testing (FDR Q value ≤ 0.05). For incident disease associations, we also observed that OMICmAge had the highest HRs for type 2 diabetes (HR = 2.68, $P = 6.14 \times 10^{-4}$), CVD (HR = 3.28, $P = 4.85 \times 10^{-6}$) and all-cause mortality (HR = 11.31, $P = 2.65 \times 10^{-23}$), which all met FDR significance (Q value ≤ 0.05). PCGrimAge, PCPhenoAge and chronological age were also FDR significant for CVD, while chronological age was the only measure that was significant for stroke (HR = 1.85, $P = 9.59 \times 10^{-4}$). No aging biomarkers were significantly associated with the incidence of depression, COPD or cancer.

We conducted similar analyses in the Generation Scotland cohort (Extended Data Fig. 6b and Supplementary Tables 14 and 15). While chronological age showed the strongest associations with all-cause mortality (HR = 5.58, $P < 1 \times 10^{-99}$), incident stroke (HR = 4.10, $P = 1.64 \times 10^{-88}$) and cancer (HR = 2.76, $P = 9.15 \times 10^{-192}$), OMICmAge generally ranked second after PCGrimAge (except cancer). OMICmAge also ranked among the top two aging biomarkers for incident type 2 diabetes (HR = 4.18, $P = 5.75 \times 10^{-25}$), CVD (HR = 4.14, $P = 2.46 \times 10^{-21}$) and COPD (HR = 1.97, $P = 1.90 \times 10^{-10}$). Notably, OMICmAge exhibited the strongest association with incident depression (HR = 3.14, $P = 1.52 \times 10^{-5}$). For prevalent disease associations, OMICmAge was also among the top two aging biomarkers by OR estimates, following PCGrimAge. However, no aging biomarkers were significantly associated with prevalent stroke after adjusting for chronological age and other covariates.

We further evaluated the prevalent disease associations in one additional cohort, the TruDiagnostic Biobank. In this cohort, PCGrimAge, OMICmAge, DNAmEMRAge and chronologic age had FDR-significant associations with type 2 diabetes, CVD, COPD and cancer. PCGrimAge had the highest OR with COPD (OR = 6.15, $P = 1.59 \times 10^{-04}$). Overall, chronological age had stronger associations with aging-related diseases in the TruDiagnostic cohort than in the other cohorts. Chronological age was the only measure that was associated with stroke (OR = 2.21, $P = 8.51 \times 10^{-15}$) and had the highest ORs with CVD (OR = 1.74, $P = 1.93 \times 10^{-172}$) and cancer (OR = 2.27, $P = 3.26 \times 10^{-146}$). OMICmAge was in the top two highest associations for type 2 diabetes (OR = 2.78, $P = 3.61 \times 10^{-13}$), depression (OR = 1.26, $P = 3.53 \times 10^{-3}$) and cancer (OR = 1.50, $P = 6.28 \times 10^{-8}$; Extended Data Fig. 6a and Supplementary Table 16).

We also calculated the area under the curve (AUC) for 5-year and 10-year survival using prediction classifiers for OMICmAge, DNAmEMRAge, PCGrimAge, chronological age and other aging biomarkers in both MGB and Generation Scotland cohorts (Fig. 5c and Extended Data Fig. 7). In the prediction models, we included age for those biomarkers in which age is not included as a feature (all except OMICmAge, DNAmEMRAge and PCGrimAge). In the MGB testing set, DNAmEMRAge showed the highest AUC values (5-year AUC: 0.898, OR = 10.77, $P = 1.14 \times 10^{-14}$; 10-year AUC: 0.89, OR = 7.99, $P = 2.17 \times 10^{-17}$), followed by OMICmAge with very similar values (5-year AUC: 0.892, OR = 14.83, $P = 5.25 \times 10^{-14}$; 10-year AUC: 0.873, OR = 10.42, $P = 2.53 \times 10^{-16}$). Chronological age and the other methylation clocks had AUC values lower than OMICmAge and DNAmEMRAge. In the Generation Scotland cohort, OMICmAge also ranked as the second-best aging biomarker based on AUC values for both the 5-year (AUC: 0.861; OR = 3.86, $P = 1.58 \times 10^{-12}$) and 10-year (AUC: 0.859; OR = 4.13, $P = 3.30 \times 10^{-29}$) periods, following PCGrimAge (5-year AUC: 0.870, OR = 8.13, $P = 9.03 \times 10^{-15}$; 10-year AUC: 0.866, OR = 8.08, $P = 2.30 \times 10^{-31}$).

Finally, we evaluated the association between OMICmAge and lifestyle factors in both MGB and TruDiagnostic Biobank cohorts with varying lifestyle information, using the FDR to identify significance after adjusting for multiple testing (FDR Q value ≤ 0.05; Fig. 6 and Supplementary Tables 17 and 18). In all the cohorts, we observed significant negative associations with female sex, education level and exercise per week. We also observed consistent significant positive associations with Black race, obesity and tobacco smoking across all cohorts. While we observed a significant positive association with being underweight in the MGB-ABC cohort, this is likely an indication of illness among individuals with low body weight that is present in the MGB Biobank and not observed in the other cohorts. This relationship has been previously reported in epidemiological studies and in the proteomics-aging clock paper developed by Oh et al.[17]. Finally, occasional recreational drug use was significantly associated with higher OMICmAge, while antioxidants and omega-3 fish oil intake were significantly associated with a lower biological age in the TruDiagnostic cohort.

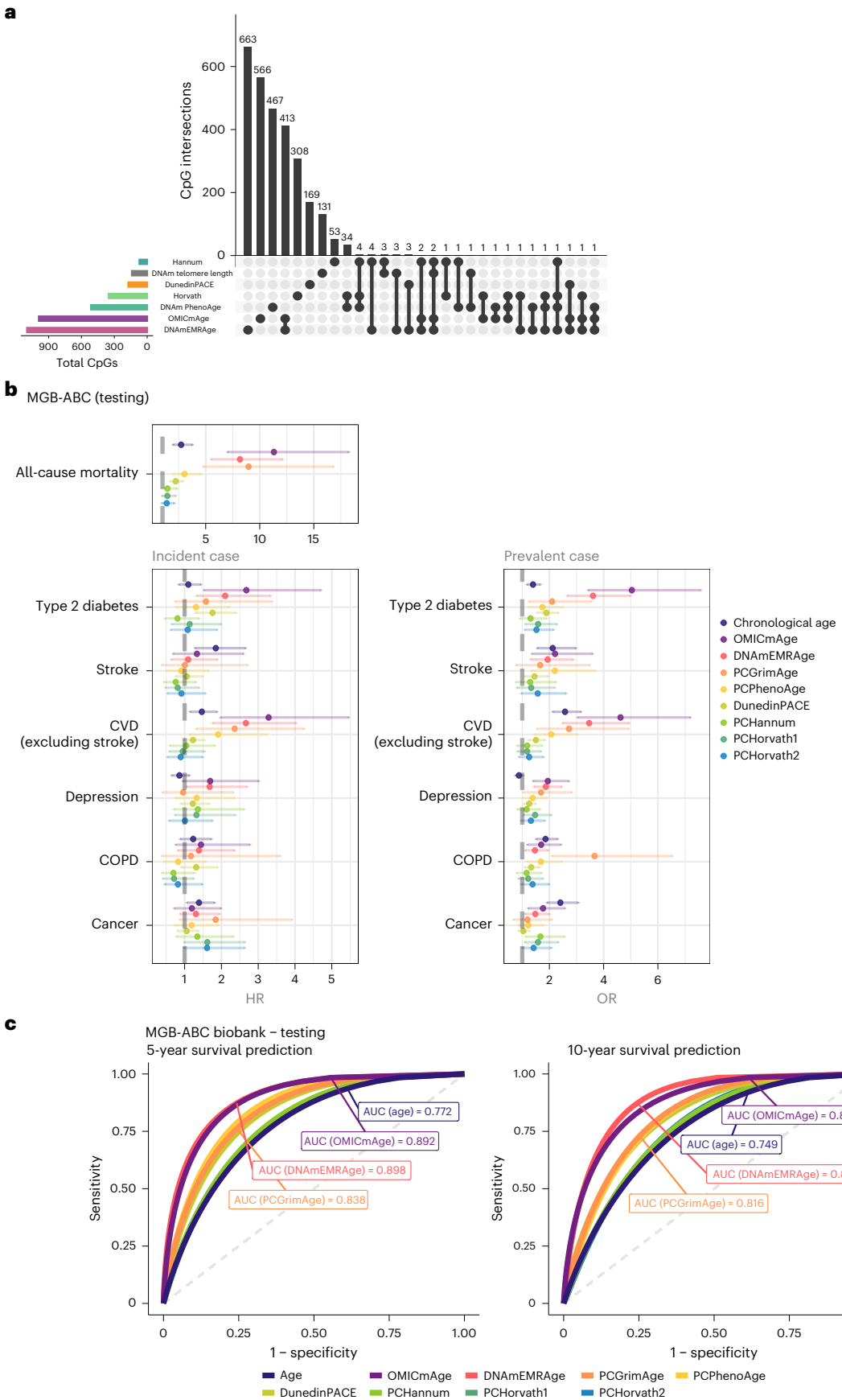
Discussion

The objective of this study was to develop a clinically relevant aging biomarker that can be implemented into the current electronic infrastructure and an analogous DNAm biomarker informed by multi-omic data. We did this through the generation of EMRAge, DNAmEMRAge and OMICmAge. The motivating premise is that biological aging is a multifactorial process involving complex interactions of cellular and biochemical processes that is best understood via multiple omic profiles. To date, aging biomarkers (also known as 'clocks') have been generated predominantly with singular omic data types[3–13]. While this approach most often generates highly predictive aging biomarkers, without additional molecular data the overall biological understanding is limited. This leaves a gap between predictive accuracy and driving physiological mechanisms.

A major goal for aging biomarkers is to implement these measures into clinical care and improve overall health. We used common clinical laboratory measures in ~30,000 individuals from the MGB Biobank to develop EMRAge. The use of readily available EMR data suggests that EMRAge can be broadly recapitulated across multiple EMR systems. While prior aging biomarkers were developed using either clinical data or mortality prediction models[31–33], EMRAge reflects a hybrid aging biomarker that distills health status and mortality into a single aging biomarker. EMRAge is highly reproducible, has strong associations with incident and prevalent chronic disease outcomes, and is an accurate predictor of mortality risk that outperforms chronologic age and PhenoAge in both our discovery cohort and a large independent cohort. The broad clinical relevance and ease of large-scale implementation highlight the translational potential of EMRAge. Further assessment in diverse populations and across different EMR systems will characterize its generalizability.

Fig. 5 | Comparison of OMICmAge and DNAmEMRAge to previously established aging biomarkers. a, Intersection of predictive CpG sites included in the previously published epigenetic clocks, DNAmEMRAge and OMICmAge. The horizontal bars represent the total number of CpG sites included in each epigenetic aging biomarker. The vertical bars represent the number of unique or shared CpG sites between clocks. PhenoAge refers to the DNAm version. **b**, Horizontal error bar plot of ORs/HRs of each methylation biomarker and chronological age to aging-related diseases in the testing set of the MGB-ABC cohort. The ratios and 95% CIs are based on a 1-s.d. change around the estimated mean values from the statistical models. The number of cases (n) and sample sizes (N) for each phenotype are as follows. Incident cases: all-cause mortality, $N = 662$ ($n = 83$); stroke, $N = 616$ ($n = 45$); type 2 diabetes, $N = 428$ ($n = 52$); COPD, $N = 485$ ($n = 39$); depression, $N = 361$ ($n = 52$); CVD (excluding stroke), $N = 287$ ($n = 82$); cancer, $N = 524$ ($n = 68$). Prevalent cases: The total sample size for prevalent diseases is 689, with the following case numbers; stroke, $n = 73$; type 2 diabetes, $n = 261$; COPD, $n = 204$; depression, $n = 328$; CVD (excluding stroke), $n = 402$; cancer, $n = 162$. **c**, Receiver operating characteristic curves comparing 5-year and 10-year survival predictions based on DNAm aging biomarkers and chronological age within the MGB-ABC cohort. Individual lines denote specific biomarkers.

https://doi.org/10.1038/s43587-026-01073-7



https://doi.org/10.1038/s43587-026-01073-7

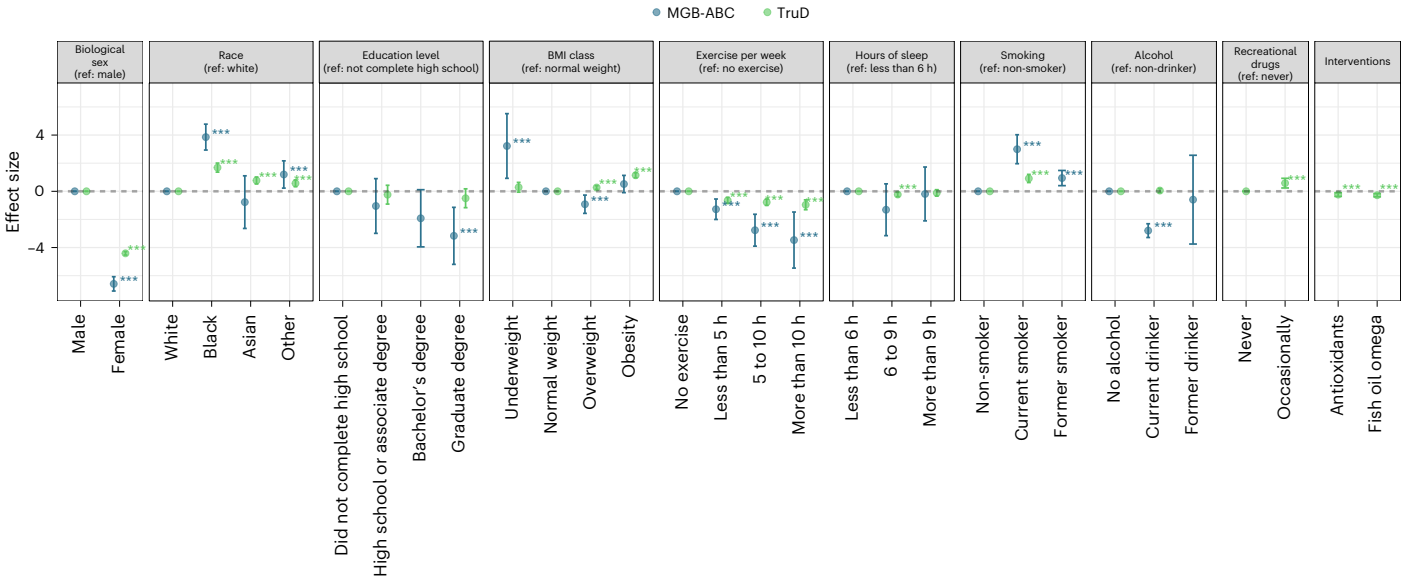


Fig. 6 | Forest plot for the representation of the lifestyle factors associated with OMICmAge in the MGB-ABC and TruDiagnostic Biobank. The effect size and 95% CIs, derived from the multivariate linear regression model, are shown as error bars for each lifestyle factor. All the associations are adjusted by chronological age, biological sex, ethnicity, BMI, tobacco use and alcohol consumption, when appropriate. The sample sizes (*N*) for each factor level across the MGB-ABC (N_1) and TruD (N_2) cohorts are presented below. FDR values are shown in parentheses for factor levels that are not the reference group (ref.). Biological sex: male (ref.), $N_1 = 1{,}158$, $N_2 = 8{,}361$; female, $N_1 = 2{,}064$ (FDR = 5.51×10^{-128}), $N_2 = 5{,}852$ (FDR < 2.22×10^{-308}). Race: white (ref.), $N_1 = 2{,}615$, $N_2 = {}_1 2{,}114$; Black, $N_1 = 241$ (FDR = ${}_1.75 \times 10^{-15}$), $N_2 = 454$ (FDR = 1.24×10^{-23}); Asian, $N_1 = 53$ (FDR = 0.473), $N_2 = 787$ (FDR = 7.28×10^{-9}); other, $N_1 = 237$ (FDR = 0.026), $N_2 = 858$ (FDR = 1.00×10^{-5}). Education Level: did not complete high school (ref.), $N_1 = 46$, $N_2 = 1{,}0$; high school or associate degree, $N_1 = 516$ (FDR = 0.343), $N_2 = 2{,}907$ (FDR = 0.610); bachelor's degree, $N_1 = 401$ (FDR = 0.097); graduate degree, $N_1 = 493$ (FDR = 0.005), $N_2 = 2{,}813$ (FDR = 0.225). BMI class: underweight, $N_1 = 35$ (FDR = 0.011), $N_2 = 396$ (FDR = 0.170); normal weight (ref.), $N_1 = 763$, $N_2 = 6{,}960$; overweight, $N_1 = 971$ (FDR = 0.011), $N_2 = 4{,}988$ (FDR = 1.88×10^{-4}); obesity, $N_1 = 1{,}355$ (FDR = 0.140), $N_2 = 1{,}869$ (FDR = 5.32×10^{-35}). Exercise per week: no exercise (ref.), $N_1 = 444$, $N_2 = {}_1{,}148$; less than 5 h, $N_1 = 737$ (FDR = 0.002), $N_2 = 7{,}477$ (FDR = 1.47×10^{-8}); 5 to 10 h, $N_1 = 155$ (FDR = 6.41×10^{-6}), $N_2 = 4{,}927$ (FDR = 6.84×10^{-11}); more than 10 h, $N_1 = 40$ (FDR = 0.002), $N_2 = 552$ (FDR = 3.29×10^{-7}). Hours of sleep: less than 6 h (ref.), $N_1 = 47$, $N_2 = 2{,}4{}_1 4$; 6 to 9 h, $N_1 = 805$ (FDR = 0.209), $N_2 = 10{,}358$ (FDR = 0.008); more than 9 h, $N_1 = 299$ (FDR = 0.847), $N_2 = 1{,}388$ (FDR = 0.462). Smoking: non-smoker (ref.), $N_1 = 2{,}052$, $N_2 = {}_1 3{,}652$; current smoker, $N_1 = 192$ (FDR = 5.13×10^{-8}), $N_2 = 561$ (FDR = 4.25×10^{-9}); former smoker, $N_1 = 978$ (FDR = 0.002). Alcohol: no alcohol (ref.), $N_1 = 1{,}609$, $N_2 = 2{,}739$; current drinker, $N_1 = 1{,}521$ (FDR = 1.20×10^{-27}), $N_2 = 11{,}474$ (FDR = 0.610); former drinker, $N_1 = 21$ (FDR = 0.769). Recreational drugs: never (ref.), $N_2 = 494$; occasionally, $N_2 = 2{,}005$ (FDR = 0.002). Interventions: antioxidants, $N_2 = 3{,}905$ (FDR = 8.4×10^{-4}); fish oil omega, $N_2 = 7085$ (FDR = 5.25×10^{-6}).

Because EMRAge alone does not resolve specific physiology, we used machine learning to predict EMRAge with DNAm and multiple omics, generating DNAmEMRAge and OMICmAge, respectively. Both aging biomarkers had excellent accuracy for 5-year and 10-year mortality risk, and DNAmEMRAge maintained the highest accuracy in one of the validation cohorts. We also demonstrated that both DNAmEMRAge and OMICmAge have strong associations with aging-related health outcomes, including CVD, stroke, type 2 diabetes, COPD, depression and mortality. Notably, among studied DNAm-based aging biomarkers, OMICmAge often exhibits either the strongest (for example, for depression) or second strongest (for example, for type 2 diabetes) association with prevalent or incident aging-related morbidities across cohorts. These patterns hold across cohorts with differing health profiles and ascertainment. The consistency in the association findings suggests that DNAmEMRAge and OMICmAge are reliable and broadly applicable across a diverse range of cohort characteristics. This argument for OMICmAge is further supported by previous studies[34–36] indicating that a significant portion of the signal captured by epigenetic aging biomarkers stems from the accumulation of stochastic variation over time. In essence, as an aging biomarker becomes more predictive (that is, higher correlation) of chronological age, it increasingly reflects a pure stochastic process, thereby demonstrating less biological relevance. Given that OMICmAge exhibited a moderate correlation with chronological age in both MGB and TruDiagnostic Biobank cohorts, its broad applicability is largely justified by its substantial deviation from purely stochastic accumulation. This suggests that OMICmAge reflects non-stochastic physiological processes related to aging.

In addition, consistent reproducibility has previously been an issue with epigenetic biomarkers, which has traditionally only been improved through the inclusion of summary features such as PCs[30,37]. With OMICmAge, we observed high ICCs, demonstrating strong reproducibility. One ongoing challenge with multi-omic aging biomarkers is the complexity of integrating different omic data together and the subsequent interpretation of the findings. Moreover, multi-omic clocks are often impractical due to high costs and logistics. The approach we used to develop OMICmAge has the advantage of estimating metabolites, proteins and clinical data (via EBPs) while distilling this into a single DNAm-based aging algorithm[29,31,38]. Using DNAm as the primary metric was selected for its stability and cost-effectiveness. In this sense, OMICmAge reflects aging processes on multiple levels of systems biology while only necessitating DNAm in its calculation.

An advantage of OMICmAge and the development with concurrent multi-omic data is the ability to further elucidate the biological mechanisms associated with this biomarker. Although the epigenome is central to aging, functions implied by specific methylation perturbations are often unclear, limiting clinical interpretability when accelerated aging arises from heterogeneous mechanisms. In contrast to the epigenome, proteins reflect a broad range of aging-related biology, including immune function and inflammatory processes that are often well understood and have clear clinical implications for treatment and/or modification. Changes in oxidative stress, hormones and lipid profiles are just a few examples of the metabolic processes reflected via the metabolome that represent specific biology relevant for aging processes[37]. When developing OMICmAge, we included protein and

metabolite EBPs that were correlated with EMRAge to improve the likelihood of retaining physiologically relevant measures. Retained EBPs included albumin[38] and the androgenic steroid androsterone sulfate. The algorithm for OMICmAge also retained protein and metabolite EBPs with a less well-understood relationship with aging, such as ribitol, which has been identified as a metabolite predictive of mortality but has very little mechanistic information[39]. It is important to highlight that not all retained features are causal, nor do they necessarily have the strongest overall associations with OMICmAge. Follow-up functional work and/or causal modeling is necessary to infer any potential causal links between these proteins/metabolites and EMRAge.

There are several limitations that need to be addressed in future work. First, EMRAge was developed using EMR data and is, therefore, primarily tailored for clinical data. The major advantage of this is that this measure uses real-world data and can be recapitulated in several EMR systems. However, real-world data are not systematically collected and so missingness is universal. To assess robustness, we calculated the median values in different time increments surrounding the time point when EMRAge is being estimated and found that the EMRAge estimates had near-perfect pairwise correlations among reconstructed EMRAge estimates. Furthermore, while EMRAge demonstrated superior performance over PhenoAge in predicting the risk of all-cause mortality using the All of Us Research Program data, EMRAge did not consistently outperform PhenoAge for other incident aging-related diseases, likely influenced by shorter follow-up periods (median: 3 years versus 5.5 years). Future work should test EMRAge across diverse populations and EMR systems to confirm validity and generalizability. A major advantage of OMICmAge is the incorporation of proteins, metabolites and clinical EBPs; however, more work is necessary to further improve the accuracy and precision of EBPs. Targeted, quantitative protein and metabolites assays will improve EBP accuracy and reflect the actual clinical levels. There is also room to expand upon the EBPs that were included into the feature space, both with additional metabolites/proteins and with other omics. Finally, additional validation of OMICmAge across diverse populations and with more aging biomarkers will continue to highlight potential advantages and limitations in this aging biomarker.

The present study introduces several notable steps in aging biomarkers. First, EMRAge advances the field as a hybrid aging biomarker that integrates clinical health data and mortality risk into a robust measure readily scalable across EMRs. Building on this foundation, we established DNAmEMRAge and OMICmAge, epigenetic aging algorithms that extend predictive accuracy and mechanistic insight. OMICmAge in particular integrates DNAm with proteomic, metabolomic and clinical data through EBPs, yet it remains measurable from DNAm alone. This systems-biology framework unifies multiple biological levels into a single readout of aging, enabling a more comprehensive and interpretable view than prior clocks. Together, these tools establish a clinically practical and biologically grounded platform for assessing biological age. Ongoing validation across diverse populations will extend their translational reach, with the potential to transform both research and clinical practice in aging and aging-related diseases.

Methods

Discovery cohort

MGB Biobank. The MGB Biobank is a large biorepository that provides access to research data and approximately 130,000 high-quality banked samples (plasma, serum and DNA) from >100,000 consented individuals enrolled in the MGB system[40]. Written informed consent was obtained from all participants upon enrollment in the biobank. Participants were linkable to EMR data spanning their MGB medical histories and to surveys on lifestyle, environment and family history. Plasma donors initially totaled 60,371 from the MGB Biobank; 124 were excluded for being <18 years old at collection. Among remaining adults, vital status for 59,213 was verified (alive/deceased) with death dates recorded as of 28 July 2022. Another 28,329 participants were excluded for missing phenotype data (Extended Data Fig. 1).

MGB-ABC. The MGB-ABC comprises 3,451 randomly selected MGB Biobank participants to yield an age-, sex- and BMI-balanced sample representative of the Biobank. For selected participants, comprehensive EMRs plus metabolomic, proteomic and epigenetic data are available. Blood samples obtained during clinical care or research draws at Brigham and Women's Hospital or Massachusetts General Hospital were used for serum, plasma and DNA/genomic analyses. Typical draws collected 30–50 ml and were linked to EMR data; the Biobank also captured additional health information at collection. Questionnaires were administered electronically or on paper and took approximately 10–15 min to complete. Items covered family history, lifestyle and environmental factors. Confidentiality and data security were prioritized: no personally identifiable information was collected, identities were protected, and survey data were encrypted.

The Phenotype Discovery Center of MGB integrates various data sources, including the Research Patient Data Registry, health information surveys and genotype results, into the Biobank Portal. This portal combines specimen data with EMR data, creating a comprehensive SQL Server database with a user-friendly web-based application[40]. Researchers can perform queries, visualize longitudinal data with timestamps, use established algorithms to define phenotypes, utilize automated natural language processing tools for analyzing EMR data using the Informatics for Integrating Biology and the Bedside (i2b2) tool kit[41] and request samples from cases and controls. Biobank Portal data include narrative clinical notes; text reports (cardiology, pathology, radiology, operative, discharge summaries); codified elements (demographics, diagnoses, procedures, labs, medications); and patient-reported exposures and family history from surveys. Additional measures (for example, lung function) were extracted using an in-house natural language processing-based algorithm.

Metabolomic profiling. Untargeted global plasma metabolomics profiling was generated by Metabolon. Coefficients of variation were measured in blinded quality-control (QC) samples randomly distributed among study samples. Batch variation was controlled for in the analysis. Sample preparation and global metabolomics profiling was performed according to methods described previously[42]. Metabolomic profiling was performed using four LC–MS methods that measure complementary sets of metabolite classes[43]: (1) amines and polar metabolites that ionize in the positive ion mode; (2) central metabolites and polar metabolites that ionize in the negative ion mode; (3) polar and nonpolar lipids; (4) free fatty acids, bile acids and metabolites of intermediate polarity. All reagents and columns for this project will be purchased in bulk from a single lot, and all instruments will be calibrated for mass resolution and mass accuracy daily[44].

Metabolite peaks were quantified by the AUC. Raw area counts for each metabolite in each sample were normalized to the run-day median to correct inter-day instrument tuning differences, setting each run's median to 1.0. Metabolites were identified by automated comparison of ion features to a ~8,000-entry reference library of chemical standards that includes retention time, molecular weight (*m/z*), preferred adducts, in-source fragments and associated mass spectrometry spectra, with visual QC using software developed at Metabolon[44]. Known chemical entities were identified by comparison to library entries of purified standards. Recurrent, structurally unnamed biochemicals generated additional spectral entries; these may be resolved upon acquisition of matching purified standards or by classical structural analysis. QC and data processing used an in-house method[45–47]. Metabolite features with signal-to-noise ratio < 10 or with undetectable/missing values in >10% of samples were excluded. Remaining missing values were imputed as half the minimum peak intensity for that feature across the cohort. Features with a pooled-sample coefficient

of variation > 25% were removed to ensure technical reproducibility. Analyses used LC–MS peak areas; values were subjected to log transformation (approximate normality, variance stabilization) and Pareto scaling to harmonize measurement scales. After QC, 1,459 metabolites from 1,986 samples remained for analysis.

Methylation profiling. DNAm data were generated with the Illumina Infinium MethylationEPIC 850 K BeadChip (>850,000 sites) covering CpG islands, non-CpG and differentially methylated sites, FANTOM5 enhancers, ENCODE open chromatin and transcription-factor binding, and microRNA promoters. Biobanked samples were stored at −80 °C and shipped to TruDiagnostic for extraction and preprocessing. From whole blood, 500 ng DNA was extracted and bisulfite-converted using the Zymo Research EZ DNA Methylation kit per the manufacturer's instructions. Converted DNA was randomly assigned to Infinium HumanMethylationEPIC chip wells. Laboratory preprocessing comprised DNA amplification, hybridization to the EPIC array, staining/washing and imaging on the Illumina iScan SQ to generate raw intensities.

Raw methylation data for the MGB Biobank were processed using the 'minfi' pipeline[48], and low-quality samples were identified using the qcfilter() function from the ENmix package[49], using default parameters. Overall, a total of 4,803 samples passed the quality assurance/QC ($P < 0.05$) and were deemed to be high-quality samples. In addition, we removed low-quality probes ($P < 0.05$ out-of-band) that were identified among the samples. This process retained 721,802 among 866,239 probes that were high quality and indicated that a large portion of the methylation data were of high quality. A combinatorial normalization processing using the Funnorm procedure ('minfi' package), followed by the RCP method (ENmix package) was performed to minimize sample-to-sample variation as noted in Foox et al.[50].

Proteomic profiling. We used the Seer proteomic platform to identify proteins and peptides related to chronological and biological aging. This uses nanoparticles with different binding capabilities to isolate and extract peptides and proteins via corona covalent attachment to the surface, paired with LC–MS/MS, and thus enables detection of low-abundance peptides and proteins.

Relative protein levels were quantified in 2,000 samples (1,600 MGB-ABC; 400 process controls) using the Proteograph Product Suite (Seer) with LC–MS. Samples were incubated with five proprietary nanoparticles on the Seer SP100 Proteograph to form protein coronas enabling physicochemical capture. Proteins were digested by trypsin, and relative levels quantified by the default data-independent acquisition method in Proteograph Analysis Software. To address peptide–protein ambiguity and improve quantification, peptides were aggregated into protein groups, then preprocessed with control-based normalization and outlier detection. This yielded estimates for 28,490 peptides across blood samples (mean 15,239) and 10,265 in controls (mean 4,281). Peptides were consolidated into 3,695 protein groups in MGB-ABC samples (mean 2,587) and 1,360 in plate controls. Following the signal drift and batch effect correction via the QC-robust spline correction algorithm[51], we applied $\log_{10}$ transformation, Pareto scaling and k-nearest-neighbor imputation based on current guidelines[48]. Stringent filters, including 80% protein presence, relative standard deviation of quality control (RSD-qc) < 0.20%, and the D-ratio comparing technical variability to biological variability < 0.70, were utilized to reinforce data validity and reliability[52]. The final processed dataset consists of 2,805 nonunique, or 536 unique, protein groups, across 1,789 samples, in which the majority of samples ($N = 1,475$) matched to methylation data.

Definition of aging-related diseases. We utilized International Classification of Diseases (ICD)-9/ICD-10 codes to identify aging-related diseases, including type 2 diabetes, COPD, depression, cancer, stroke and other CVDs, as detailed in Supplementary Table 19. We also utilized SNOMED codes to identify the same set of diseases for the All of Us cohort, as detailed in Supplementary Table 20.

Validation cohort

All of Us cohort. The All of Us Research Program, an initiative launched by the National Institutes of Health (NIH) in 2018, represents a national collaborative effort to aggregate genetic, lifestyle, environmental and EMR data from one million participants[53]. All the participants provided a written consent form at the time of enrollment. Given the diverse data sources utilized by All of Us, including EMR data from healthcare facilities, participant surveys and self-reported measurements, the program implemented the Observational Medical Outcomes Partnership Common Data Model[54] to store and standardize this heterogeneous data, which were coded using the SNOMED CT dictionary. To facilitate the efficient retrieval of diagnosis records, we mapped the curated ICD-9 and ICD-10 codes to corresponding SNOMED CT codes, as detailed in Supplementary Tables 20 and 21.

On 4 February 2025, All of Us released the latest version of its Curated Data Repository (CDR v8), encompassing participant data up to a cutoff date of 1 October 2023. Within CDR v8, 389,379 adult participants had both EMR and lifestyle survey data available. Following that, we extracted demographic information, lifestyle factors, diagnosis records and laboratory test results necessary for the calculation of EMRAge and PhenoAge. To address missing laboratory values and ensure a sufficient sample size for association analyses, we imputed missing values with the median of all measurements recorded within one year of each participant's enrollment date. This imputation process resulted in a final cohort of 10,769 participants with complete data from the All of Us Research Program.

TruDiagnostic Biobank cohort. The TruDiagnostic Biobank included 14,698 individuals who underwent the commercial TruDiagnostic TruAge test with DNAm profiling. Participants, recruited from October 2020 to April 2023, were predominantly based in the United States and generally healthier than MGB Biobank participants, likely reflecting proactive health interest and willingness to pay for testing. Most samples were obtained under healthcare provider guidance, and <5% were via direct-to-consumer testing. This recruitment likely introduces self-selection toward preventive care and fewer comorbidities. At enrollment, participants completed surveys covering personal information, medical, social, lifestyle and family history. The study was approved by the Institute for Regenerative and Cellular Medicine Institutional Review Board, and all participants provided written informed consent.

Peripheral blood was collected via lancet/capillary, placed in lysis buffer and DNA extracted. Bisulfite conversion of 500 ng DNA was performed with the Zymo Research EZ DNA Methylation kit per the manufacturer's instructions. Converted DNA was randomly assigned to Infinium HumanMethylationEPIC BeadChip wells, then amplified, hybridized, stained/washed and imaged on an Illumina iScan SQ to generate raw intensities.

TruDiagnostic methylation data were preprocessed using the MGB-ABC pipeline, with normalization adapted for computational constraints. In total, 14,213 individuals (96.7% of originals) passed quality assurance/QC ($P < 0.05$). To retain CpGs needed for clock calculation, no probes were removed. We applied normal-exponential out-of-band (Noob) normalization using minfi's preprocessNoob function. Finally, we used a 12-cell immune deconvolution method to estimate cell-type proportions[55–57].

Generation Scotland. Generation Scotland is a Scottish, family-based cohort study with over 24,000 volunteers, aged 17–99, stemming from >5,500 families[58]. The majority of volunteers provided blood samples at a baseline clinic between 2006 and 2011 in addition to completing health and lifestyle questionnaires and giving consent for data linkage

to their electronic health records. 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. This study was performed in accordance with the Helsinki declaration.

DNAm data have been profiled using the Illumina EPICv1 array. QC details have been described previously[59]. Briefly, samples were assessed in four sets yielding data for 18,869 individuals after QC (*N*-set1 = 5,087, *N*-set2 = 459, *N*-set3 = 4,450, *N*-set4 = 8,873). Here, after the removal of 11 individuals who subsequently withdrew consent, we had data for 18,858 volunteers. Secondary care linkage was available for 99% of Generation Scotland volunteers (*N*-analysis = 18,672). OMICmAge and the other epigenetic biomarkers were estimated as described for the MGB-ABC cohort.

Event status and age at event for six disease outcomes and all-cause mortality were determined via linkage to electronic health records. The secondary (ICD) and primary care (READ) codes used to define each outcome are listed in Supplementary Table 22. Secondary care linkage was available for 99% of Generation Scotland volunteers. While all volunteers provided consent for linkage to primary care records, currently these are only available for ~40% of volunteers due to consent constraints with the data holders (individual GP surgeries). The latest date of linkage for primary and secondary care records was October 2023, which was set at the censoring date for the time-to-event analyses. Mortality records up to October 2023 were obtained via linkage to the National Records of Scotland.

Statistical analysis

Development of EMRAge. From 60,370 MGB Biobank participants with plasma samples, we extracted 28 clinical phenotypes (Supplementary Table 2 and Extended Data Fig. 1) spanning sociodemographics, clinical chemistries and behaviors (for example, smoking, alcohol). To address missingness and instrument variation, all phenotypes except height and age were subjected to median imputation using observations within 5 years of the first plasma collection. Categorical variables (for example, sex, race, Charlson Comorbidity Index, smoking status, alcohol status) were one-hot encoded; numerical variables were standardized, yielding 43 clinical variables from 28 phenotypes. Participants with complete variables (n = 29,222) entered a two-step selection for EMRAge. First, a LASSO Cox model predicting time to death (Harrell's C) selected 30 variables. Second, each variable's association with time to death was tested in a Cox proportional-hazards model; variables with adjusted $P \leq 0.05$ were retained, leaving 19. Re-adding 2,042 participants with complete data for these 19 variables yielded a final cohort of 31,264 participants and 19 variables for EMRAge development.

We split 31,264 samples by 70:30 into training (N = 21,885) and testing (N = 9,379) sets. A Cox proportional-hazards model was fitted on the training set to estimate coefficients for the 19 variables; variables were not scaled to preserve generalizability. For each individual, the linear predictor Xβ provided a risk estimate, which we transformed to EMRAge (matching the mean and variance of chronological age): EMRAge = 9.70296 × $X\beta_{train}$ + 51.68254. We then evaluated the Pearson correlation of EMRAge to chronological age in the training and testing sets.

Because EMRAge predictors were selected on imputed data, we tested their robustness by retraining at four time points (1 January 2008, 2010, 2012 and 2014). For each time point, selected clinical variables were replaced with medians from a 1-year window, ensuring no overlap of data scans. To enable fair comparison, Charlson Comorbidity Index indicators were excluded as relatively time insensitive. This yielded four estimating equations (same predictors, different data). We applied these to participants with complete predictors within a 1-year window centered on 1 January 2016 (N = 11,673); we computed and assessed Pearson correlations among the four EMRAge estimates.

We assessed associations of EMRAge with incident and prevalent aging-related outcomes (all-cause mortality, stroke, type 2 diabetes, COPD, depression, other CVD, any cancer). Cases required ≥2 relevant ICD-9/ICD-10 codes with first and last codes ≥1 day apart. Incident risks were modeled with Cox proportional hazards adjusting for age, sex, race, BMI, smoking and alcohol; prevalent morbidities were modeled with logistic regression using the same covariates. Effect sizes are reported per s.d. increase in EMRAge.

Construction of EMRAge and PhenoAge in All of Us cohort. Using the CDR v8 dataset, we identified all participants with both EMR and survey data. We then utilized the Observational Medical Outcomes Partnership standard concept IDs to retrieve clinical measurements for the following biomarkers: BMI ('3038553'), diastolic blood pressure ('3012888'), albumin ('3024561'), alanine aminotransferase (ALT, '3006923', '46235106'), alkaline phosphatase (ALP, '3001110', '3035995'), aspartate aminotransferase (AST, '3013721'), eosinophil counts ('3013115', '3028615', '3009932'), red cell distribution width (RDW, '3002888', '3015182', '3002385', '3019897'), glucose ('3037110', '3000483', '3004501'), hematocrit ('3009542', '3023314'), neutrophil counts ('3017732', '3013650', '3017501'), platelet counts ('3007461', '3024929'), triglycerides ('3022038', '3022192') and blood urea nitrogen ('3004295', '3013682'). Additionally, we mapped ICD-9 and ICD-10 codes to SNOMED CT codes, as detailed in Supplementary Table 21, for the calculation of the Charlson Comorbidity Index. Furthermore, we extracted smoking-related lifestyle survey data to categorize participants into smokers and non-smokers/former smokers. A similar retrieval method was used to extract additional clinical measurements for C-reactive protein ('3020460', '3010156'), creatinine ('3016723'), leukocyte counts (white blood cell count, '3010813', '3000905'), lymphocyte percentage ('3002030', '3037511', '3038058'), and mean corpuscular volume ('3024731', '3023599') for the construction of PhenoAge.

Comparison of EMRAge and PhenoAge. We compared EMRAge and PhenoAge regarding their associations with the incidence and prevalence of aging-related diseases. This was done by applying the same regression models described above, using either PhenoAge or EMRAge as the main predictor. We calculated PhenoAge for our MGB Biobank cohort and All of Us cohort using the established tool kit, as initially proposed by Levine et al.[19] and developed by Belsky and Kwon[60]. Although PhenoAge is derived from eight clinical lab metrics, one specific parameter, C-reactive protein, is not frequently ordered in routine clinical settings. To maximize sample retention, we used the same strategy to retain the median value of observations over a 5-year (1-year for All of Us cohort) window centered on the plasma collection date. Following this imputation, our sub-cohort of the MGB Biobank cohort consisted of 17,252 participants, with 12,081 samples in the training set and 5,171 in the testing set, while the All of Us cohort finally consisted of 10,769 samples.

Additionally, we also conducted a comparative analysis by including both EMRAge and PhenoAge into the same regression model without additional covariates. HRs and ORs per s.d. were estimated for each aging-related health outcome in both the MGB training and testing sets, as well as in the All of Us cohort.

Development of DNAmEMRAge. After developing the EMRAge measure, we next created a DNAm surrogate predictor of EMRAge using matched EPIC array data (DNAmEMRAge). To this end, we used the MGB-ABC cohort, which is a subset of the MGB Biobank that was created with the aim of possessing a proportionate aging biobank population. DNAm was generated from a total of 4,803 samples using the EPICv1 array. To allow for training, samples were then selected for having EMRAge quantified and the availability of chronological age and sex information, which retained 3,451 samples. Using an 80:20 train–test split, the glmnet R package[61] (version 4.1-8) was used to train a Gaussian penalized regression model using an alpha parameter of 0.1. Based

on a previous paper that demonstrates the benefits of increasing the number of folds in cross-validation[62], we used 25-fold to identify an optimal lambda based on the alpha parameter. Sex was classified as Gender_M (males) and Gender_F (females) using one-hot encoding and was included as penalized features along with chronological age and the relative cell proportions of 12 immune cell types. All CpGs and the covariates mentioned were included as penalized features. Those features that showed a nonzero coefficient were selected for the final model.

Development of EBP models. To reduce dimensionality, mitigate multicollinearity and enhance computational efficiency, we selected proteins and clinical phenotypes exhibiting a Pearson correlation coefficient greater than 0.1 with EMRAge. For metabolites, we used hierarchical clustering to group similar metabolites into distinct clusters. This approach leverages the observation that metabolites within the same cluster are highly intercorrelated, whereas metabolites between different clusters show only moderate-to-low correlation. Subsequently, from each identified cluster, we selected the metabolite demonstrating the strongest correlation with EMRAge.

For each significant clinical, metabolite and protein-group outcome, we fit Gaussian penalized regression (glmnet) with 25-fold cross-validation to select λ. We set $\alpha = 0.1$ to maximize CpG inclusion, except for smoking pack-years, total bilirubin and total cholesterol, where $\alpha = 0.5$ yielded the highest predicted-versus-observed correlation. Sex (one-hot: Gender_M, Gender_F) and chronological age were included as penalized covariates. Features with nonzero coefficients define each EBP. EBPs showing Pearson $\rho > 0.20$ with $P < 0.05$ were retained as inputs for OMICmAge.

Development of OMICmAge. OMICmAge was also developed using the MGB-ABC cohort ($N = 3,451$) with an 80:20 train–test split. Penalized regression (alpha = 0.1) with 25-fold cross-validation to select lambda trained a single composite model on the training set to predict EMRAge. Predictors included all QC-passed CpG sites; selected clinical, metabolite and protein EBP estimates; relative abundances of 12 immune cell subtypes; and demographics (age, sex, BMI). Sex was one-hot encoded (Gender_M, Gender_F). Features with nonzero coefficients were retained in the final multivariate model.

Comparison of OMICmAge, DNAmEMRAge and previous clocks. For comparison of OMICmAge and DNAmEMRAge clocks to previous methods of biological age prediction, we chose to analyze PCHorvath[10], PCHannum[11], PCPhenoAge[19], PCGrimAge[21] and DunedinPACE[22]. We chose their PC versions as they have much better precision while still maintaining their relationships to health outcomes[63]. To compare the CpG sites included in each model, we used the non-PC clocks because the PC models do not contain CpG sites as predictors.

We evaluated associations between each clock and incident aging-related diseases (type 2 diabetes, stroke, depression, COPD, other CVDs, any type of cancer) using Cox proportional-hazards regression, and with prevalent diseases using logistic regression. All models adjusted for age, gender, race, BMI, smoking status and alcohol drinking habits. To predict 4-year, 5-year and 10-year survival, we fit simple logistic models with a binary survival flag as the outcome and each clock as the sole predictor, then generated receiver operating characteristic curves and model AUCs. For fair comparison across biomarkers, all metrics were standardized by subtracting the mean and dividing by the s.d., making DunedinPACE performance comparable to other aging biomarkers.

Statistics and reproducibility

Programs and sample sizes. All analyses were conducted in R (version 4.3.0 or later) using standard packages including minfi, ENmix, dplyr, ggplot2, glmnet, survival, pROC and stats. The primary training and validation datasets comprised 31,264 individuals from the MGB Biobank with complete clinical data, used for training ($n = 21,885$) and testing ($n = 9,379$) of EMRAge. Independent testing was performed in 10,769 participants from the All of Us cohort. The MGB-ABC ($n = 3,451$) with clinical and methylation data was used for training ($n = 2,762$) and testing ($n = 689$) of DNAmEMRAge and OMICmAge. Additional validation datasets included Generation Scotland ($n = 18,672$) and the TruDiagnostic Biobank ($n = 14,213$ with self-reported health and lifestyle information). Technical reliability of DNAmEMRAge and OMICmAge was assessed using 30 duplicate blood samples from the MGB-ABC cohort.

Statistical models. Cox proportional-hazards models were applied to time-to-event outcomes (for example, all-cause mortality, incident cancer), whereas binomial logistic regression was used for binary outcomes (for example, cancer prevalence at plasma collection, sex). Multinomial logistic regression was applied for categorical traits with more than two classes (for example, smoking status and BMI category). Model assumptions included conditional independence and linearity for logistic regression and proportional hazards for Cox models. All input variables were standardized before modeling. Data were assumed to follow an approximately normal distribution, although this assumption was not formally tested. For technical reliability estimation, ICCs were calculated using a two-way random-effects model.

Multiple testing and blinding. To account for multiple hypothesis testing, Benjamini–Hochberg FDR correction was applied across all aging biomarkers tested per phenotype. As these were prospective observational studies, there was no blinding to the conditions of the experiments.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The data used in this study were generated from two biobank cohorts housed at MGB and TruDiagnostic, respectively. Requests for raw data, analyzed data and materials used in MGB and TruDiagnostic Biobank cohorts to generate the results in this study will be reviewed by the contact PIs (J.L.-S. or V.B.D.). Due to participant consent restrictions, individual-level data linked to electronic health records cannot be made publicly available. However, we are committed to supporting scientific collaboration and transparency wherever possible. Researchers interested in noncommercial academic use of the data, algorithms or derived biomarkers are encouraged to contact the corresponding author (J.L.-S.) or first author (V.B.D., varun@trudiagnostic.com). Upon review, a collaborative agreement or data use agreement can be established to enable:

- Access to de-identified omics data for research purposes
- Application of algorithms or generation of derived biomarker scores on external datasets
- Internal use of electronic health record-linked clinical data for validation, replication or joint analyses with external collaborators

Requests will be assessed to determine whether they are subject to intellectual property or confidentiality obligations. Access to certain raw or analyzed data may be restricted if it relates to ongoing or future patent filings, or if release could compromise the commercialization strategy of MGB or TruDiagnostic. Access may also be limited for data containing proprietary algorithms or other commercially sensitive information. We welcome such requests and aim to respond within 30 days.
Publicly available datasets referenced in this study include:

- All of Us Research Program: https://allofus.nih.gov/
- Generation Scotland: https://genscot.ed.ac.uk/
- SEER proteomics data for the MGB-ABC cohort are available under PRIDE accession PXD048709

Code availability

Code to calculate all metrics is accessible via TruDiagnostic's DNAm Analysis Software. You can request access to the software at https://www.trudiagnostic.com/softwarerequest/, at no cost for research purposes. Access is contingent upon the requestor agreeing to a standard, noncommercial software license agreement that confirms the intended use is strictly for academic, nonprofit research and that the requestor will not attempt to reverse engineer or commercially exploit the software. All R codes used to develop and validate EMRAge, DNAmEMRAge and OMICmAge are available at https://github.com/LaskySuLab/OMICmAge/.

References

1. Ferrucci, L. & Kuchel, G. A. Heterogeneity of aging: individual risk factors, mechanisms, patient priorities, and outcomes. *J. Am. Geriatr. Soc.* **69**, 610–612 (2021).
2. Jansen, R. et al. An integrative study of five biological clocks in somatic and mental health. *eLife* **10**, e59479 (2021).
3. Zhang, W. -G. et al. Select aging biomarkers based on telomere length and chronological age to build a biological age equation. *Age* **36**, 9639 (2014).
4. Cole, J. H. et al. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. *NeuroImage* **163**, 115–124 (2017).
5. Cole, J. H. et al. Brain age predicts mortality. *Mol. Psychiatry* **23**, 1385–1392 (2018).
6. Sokolova, K., Barker, G. J. & Montana, G. Convolutional neural-network-based ordinal regression for brain age prediction from MRI scans. In *Medical Imaging 2020: Image Processing* 11313 (eds Išgum, I. & Landman, B. A.) 572–579 (SPIE, 2020).
7. Goyal, M. S. et al. Persistent metabolic youth in the aging female brain. *Proc. Natl Acad. Sci. USA* **116**, 3251–3255 (2019).
8. Alpert, A. et al. A clinically meaningful metric of immune age derived from high-dimensional longitudinal monitoring. *Nat. Med.* **25**, 487–495 (2019).
9. Bocklandt, S. et al. Epigenetic predictor of age. *PLoS ONE* **6**, e14821 (2011).
10. Horvath, S. DNA methylation age of human tissues and cell types. *Genome Biol.* **14**, R115 (2013).
11. Hannum, G. et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol. Cell* **49**, 359–367 (2013).
12. van den Akker, E. B. et al. Metabolic age based on the BBMRI-NL ^{1}H-NMR metabolomics repository as biomarker of age-related disease. *Circ. Genom. Precis. Med.* **13**, 541–547 (2020).
13. Krištić, J. et al. Glycans are a novel biomarker of chronological and biological ages. *J. Gerontol. Ser. A Biol. Sci. Med. Sci.* **69**, 779–789 (2014).
14. Lehallier, B., Shokhirev, M. N., Wyss-Coray, T. & Johnson, A. A. Data mining of human plasma proteins generates a multitude of highly predictive aging clocks that reflect different aspects of aging. *Aging Cell* **19**, e13256 (2020).
15. Lehallier, B. et al. Undulating changes in human plasma proteome profiles across the lifespan. *Nat. Med.* **25**, 1843–1850 (2019).
16. Enroth, S., Enroth, S. B., Johansson, Å & Gyllensten, U. Protein profiling reveals consequences of lifestyle choices on predicted biological aging. *Sci. Rep.* **5**, 17282 (2015).
17. Oh, H. S. -H. et al. Organ aging signatures in the plasma proteome track health and disease. *Nature* **624**, 164–172 (2023).
18. Pendergrass, S. A. & Crawford, D. C. Using electronic health records to generate phenotypes for research. *Curr. Protoc. Hum. Genet.* **100**, e80 (2019).
19. Levine, M. E. et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging* **10**, 573–591 (2018).
20. Qiu, W., Chen, H., Kaeberlein, M. & Lee, S. -I. ExplaiNAble BioLogical Age (ENABL Age): an artificial intelligence framework for interpretable biological age. *Lancet Healthy Longev.* 4, e711–e723 (2023).
21. Lu, A. T. et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* **11**, 303–327 (2019).
22. Belsky, D. W. et al. DunedinPACE, a DNA methylation biomarker of the pace of aging. *eLife* **11**, e73420 (2022).
23. Schaum, N. et al. Aging hallmarks exhibit organ-specific temporal signatures. *Nature* **583**, 596–602 (2020).
24. Sehgal, R. et al. Systems Age: a single blood methylation test to quantify aging heterogeneity across 11 physiological systems. *Nat. Aging* **5**, 1880–1896 (2025).
25. Rutledge, J., Oh, H. & Wyss-Coray, T. Measuring biological age using omics data. *Nat. Rev. Genet.* **23**, 715–727 (2022).
26. Panyard, D. J., Yu, B. & Snyder, M. P. The metabolomics of human aging: advances, challenges, and opportunities. *Sci. Adv.* **8**, eadd6155 (2022).
27. Thompson, M. et al. Methylation risk scores are associated with a collection of phenotypes within electronic health record systems. *npj Genom. Med.* **7**, 50 (2022).
28. Conole, E. L. S. et al. DNA methylation and protein markers of chronic inflammation and their associations with brain and cognitive aging. *Neurology* **97**, e2340–e2352 (2021).
29. Gadd, D. A. et al. Epigenetic scores for the circulating proteome as tools for disease prediction. *eLife* **11**, e71802 (2022).
30. Principal component analysis improves reliability of epigenetic aging biomarkers. *Nat. Aging* **2**, 578–579 (2022).
31. Lu, A. T. et al. DNA methylation GrimAge version 2. *Aging* **14**, 9484–9549 (2022).
32. McGreevy, K. M. et al. DNAmFitAge: biological age indicator incorporating physical fitness. *Aging* **15**, 3904–3938 (2023).
33. Macdonald-Dunlop, E. et al. A catalogue of omics biological ageing clocks reveals substantial commonality and associations with disease risk. *Aging* **14**, 623–659 (2022).
34. Meyer, D. H. & Schumacher, B. Aging clocks based on accumulating stochastic variation. *Nat. Aging.* **4**, 871–885 (2024).
35. Tong, H. et al. Quantifying the stochastic component of epigenetic aging. *Nat. Aging.* **4**, 886–901 (2024).
36. Higgins-Chen, A. T. et al. A computational solution for bolstering reliability of epigenetic clocks: implications for clinical trials and longitudinal tracking. *Nat. Aging* **2**, 644–661 (2022).
37. Bizarri, D. et al. NMR metabolomics-guided DNA methylation mortality predictors. *eBioMedicine* **107**, 105279 (2024).
38. Hirata, T. et al. Associations of cardiovascular biomarkers and plasma albumin with exceptional survival to the highest ages. *Nat. Commun.* **11**, 3820 (2020).
39. Zhou, L. et al. Integrated proteomic and metabolomic modules identified as biomarkers of mortality in the Atherosclerosis Risk in Communities study and the African American Study of Kidney Disease and Hypertension. *Hum. Genomics* **16**, 53 (2022).
40. Castro, V. M. et al. The Mass General Brigham Biobank Portal: an i2b2-based data repository linking disparate and high-dimensional patient data to support multimodal analytics. *J. Am. Med. Inform. Assoc.* **29**, 643–651 (2022).
41. Murphy, S. N. et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). *J. Am. Med. Inform. Assoc.* **17**, 124–130 (2010).
42. Evans, A. M., DeHaven, C. D., Barrett, T., Mitchell, M. & Milgram, E. Integrated, nontargeted ultrahigh performance liquid chromatography/electrospray ionization tandem mass spectrometry platform for the identification and relative quantification of the small-molecule complement of biological systems. *Anal. Chem.* **81**, 6656–6667 (2009).

43. Sha, W. et al. Metabolomic profiling can predict which humans will develop liver dysfunction when deprived of dietary choline. *FASEB J.* **24**, 2962–2975 (2010).
44. Dehaven, C. D., Evans, A. M., Dai, H. & Lawton, K. A. Organization of GC/MS and LC/MS metabolomics data into chemical libraries. *J. Cheminform.* **2**, 9 (2010).
45. Kelly, R. S. et al. Integration of metabolomic and transcriptomic networks in pregnant women reveals biological pathways and predictive signatures associated with preeclampsia. *Metabolomics* **13**, 7 (2017).
46. Townsend, M. K. et al. Impact of pre-analytic blood sample collection factors on metabolomics. *Cancer Epidemiol. Biomarkers Prev.* **25**, 823–829 (2016).
47. Mayers, J. R. et al. Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development. *Nat. Med.* **20**, 1193–1198 (2014).
48. Aryee, M. J. et al. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).
49. Xu, Z., Niu, L., Li, L. & Taylor, J. A. ENmix: a novel background correction method for Illumina HumanMethylation450 BeadChip. *Nucleic Acids Res.* **44**, e20 (2016).
50. Foox, J. et al. The SEQC2 epigenomics quality control (EpiQC) study. *Genome Biol.* **22**, 332 (2021).
51. Kirwan, J. A., Broadhurst, D. I., Davidson, R. L. & Viant, M. R. Characterising and correcting batch variation in an automated direct infusion mass spectrometry (DIMS) metabolomics workflow. *Anal. Bioanal. Chem.* **405**, 5147–5157 (2013).
52. Broadhurst, D. et al. Guidelines and considerations for the use of system suitability and quality control samples in mass spectrometry assays applied in untargeted clinical metabolomic studies. *Metabolomics* **14**, 72 (2018).
53. All of Us Research Program Investigators; Denny, J. C. et al. The "All of Us" Research Program. *N. Engl. J. Med.* **381**, 668–676 (2019).
54. Stang, P. E. et al. Advancing the science for active surveillance: rationale and design for the Observational Medical Outcomes Partnership. *Ann. Intern. Med.* **153**, 600–606 (2010).
55. Luo, Q. et al. A meta-analysis of immune-cell fractions at high resolution reveals novel associations with common phenotypes and health outcomes. *Genome Med.* **15**, 59 (2023).
56. Zheng, S. C., Breeze, C. E., Beck, S. & Teschendorff, A. E. Identification of differentially methylated cell types in epigenome-wide association studies. *Nat. Methods* **15**, 1059–1066 (2018).
57. Faul, J. D. et al. Epigenetic-based age acceleration in a representative sample of older Americans: associations with aging-related morbidity and mortality. *Proc. Natl Acad. Sci. USA* **120**, e2215840120 (2023).
58. Milbourn, H. et al. Generation Scotland: an update on Scotland's longitudinal family health study. *BMJ Open* **14**, e084719 (2024).
59. Walker, R. M. et al. Data Resource Profile: whole blood DNA methylation resource in generation scotland (MeGS). *Int. J. Epidemiol.* **54**, dyaf091 (2025).
60. Kwon, D. & Belsky, D. W. A toolkit for quantification of biological age from blood chemistry and organ function test data: BioAge. *Geroscience* 43, 2795–2808 (2021).
61. Tay, J. K., Narasimhan, B. & Hastie, T. Elastic net regularization paths for all generalized linear models. *J. Stat. Softw.* **106**, 1–31 (2023).
62. Bernabeu, E. et al. Refining epigenetic prediction of chronological and biological age. *Genome Med.* **15**, 12 (2023).
63. Smirnov, D., Mazin, P., Osetrova, M., Stekolshchikova, E. & Khrameeva, E. The Hitchhiker's Guide to untargeted lipidomics analysis: practical guidelines. *Metabolites* **11**, 713 (2021).

Acknowledgements

R.K., K.M., Y.C., S.B. and J.A.L.S. are supported by R01HL123915 from the NIH/NHLBI. P.K. is supported by K99HL159234 from the NIH/NHLBI. R.S.K. is supported by K01HL146980 from the NIH/NHLBI. S.H.C. is supported by K01HL153941 from the NIH/NHLBI. Y.C. and J.A.L.S. are supported by R01HL141826 from the NIH/NHLBI. A.D. is supported by K01HL130629 from the NIH/NHLBI. A.D. and J.A.L.S. are supported by 1R01HL152244 from the NIH/NHLBI. M.M. and J.A.L.S. are supported by R01HL155742 from the NIH/NHLBI. M.M. is supported by R01HL139634 from the NIH/NHLBI. Q.C. and J.A.L.S. are supported by R01HL169300 from the NIH/NHLBI. J.A.L.S., S.T.W. and E.W.K. are supported by the NIH U01HG008685. V.N.G. is supported by NIH AG065403, AG064223 and by Hevolution. C.E.W. is supported by the Swedish Heart Lung Foundation (HLF 2023-0463 and HLF 2021-0519), and the Swedish Research Council (2022-00796). N.J.W.R. is supported by MR/Y010736/1 from the Medical Research Council and IF\R1\231034 from the Royal Society. E.M. is supported in part by the intramural program Metabolomics and multi-omics (ZIC TR000547) at the National Center for Advancing Translational Sciences, part of the NIH. DNAm profiling of the Generation Scotland samples was carried out by the Genetics Core Laboratory at the Edinburgh Clinical Research Facility, University of Edinburgh and was funded by the Medical Research Council UK and Wellcome Trust (104036/Z/14/Z and 220857/Z/20/Z). Core support was received from the Chief Scientist Office of the Scottish Government Health Directorates (CZD/16/6) and the Scottish Funding Council (HR03006). TruDiagnostic generated the DNAm and proteomic data. This research was supported in part by the Intramural Research Program of the NIH. The contributions of the NIH author(s) were made as part of their official duties as NIH federal employees, are in compliance with agency policy requirements, and are considered Works of the US Government. However, the findings and conclusions presented in this paper are those of the authors and do not necessarily reflect the views of the NIH or the US Department of Health and Human Services.

Author contributions

Q.C. had full access to the MGB-ABC data and verified the data integrity and accuracy of the analysis. V.B.D. had access to the omics data and the TruDiagnostic Cohort and verified the data integrity and accuracy of the analysis. Conceived and designed the study: J.L.-S., R. Smith and V.B.D. Sample selection: S.H.C., R.S.K., Y.C. and S.B. Performed analysis of physical samples from Research Patient Data Registry: Y.C., S.B., A.D., M.M., C.E.W. and E.A.M. Funding support of omics: A.D., M.M., C.E.W., E.M., J.L.-S. and R. Smith. Performed analysis of physical samples for the TruDiagnostic cohort: T.M. and H.W. Data processing and normalization: Q.C., V.B.D., K.M. and P.K. Algorithm development: Q.C., V.B.D. and J.L.-S. Statistical analysis and validation: Q.C., V.B.D. and N.C.-G. Validation analyses in the Generation Scotland cohort: R.E.M. and A.R. Validation in All of Us: Q.C. and M.A.A. All authors drafted and edited the paper.

Competing interests

This work was completed under a sponsored research agreement between Brigham Women's Hospital and TruDiagnostic. N.C.-G., L.B.D., I.G., R.D., D.P., A.T.H.-C., S.V., S.H., T.M., R. Sehgal and V.B.D. are employees of TruDiagnostic. J.L.-S. is a scientific advisor to TruDiagnostic, Precision and Ahara. R.E.M. is a scientific advisor to the Epigenetic Clock Development Foundation and Optima Partners. J.L.-S., Q.C., V.B.D. and R. Smith have filed patents on work from this publication. M.M. and V.N.G. have filed patents on measuring

cellular aging. S.T.W. receives royalties from UpToDate and is on the Board of Histolix, a digital pathology company. R.E.M. is a scientific advisor to the Epigenetic Clock Development Foundation and Optima Partners. R. Sehgal is a scientific advisor for TruDiagnostic and has received consulting fees from the company. TruDiagnostic generated the DNAm and proteomic data. Investigators from TruDiagnostic co-analyzed and co-wrote the paper as described in the contributions above. The other authors declare no competing interests.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s43587-026-01073-7.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43587-026-01073-7.

Correspondence and requests for materials should be addressed to Jessica Lasky-Su.

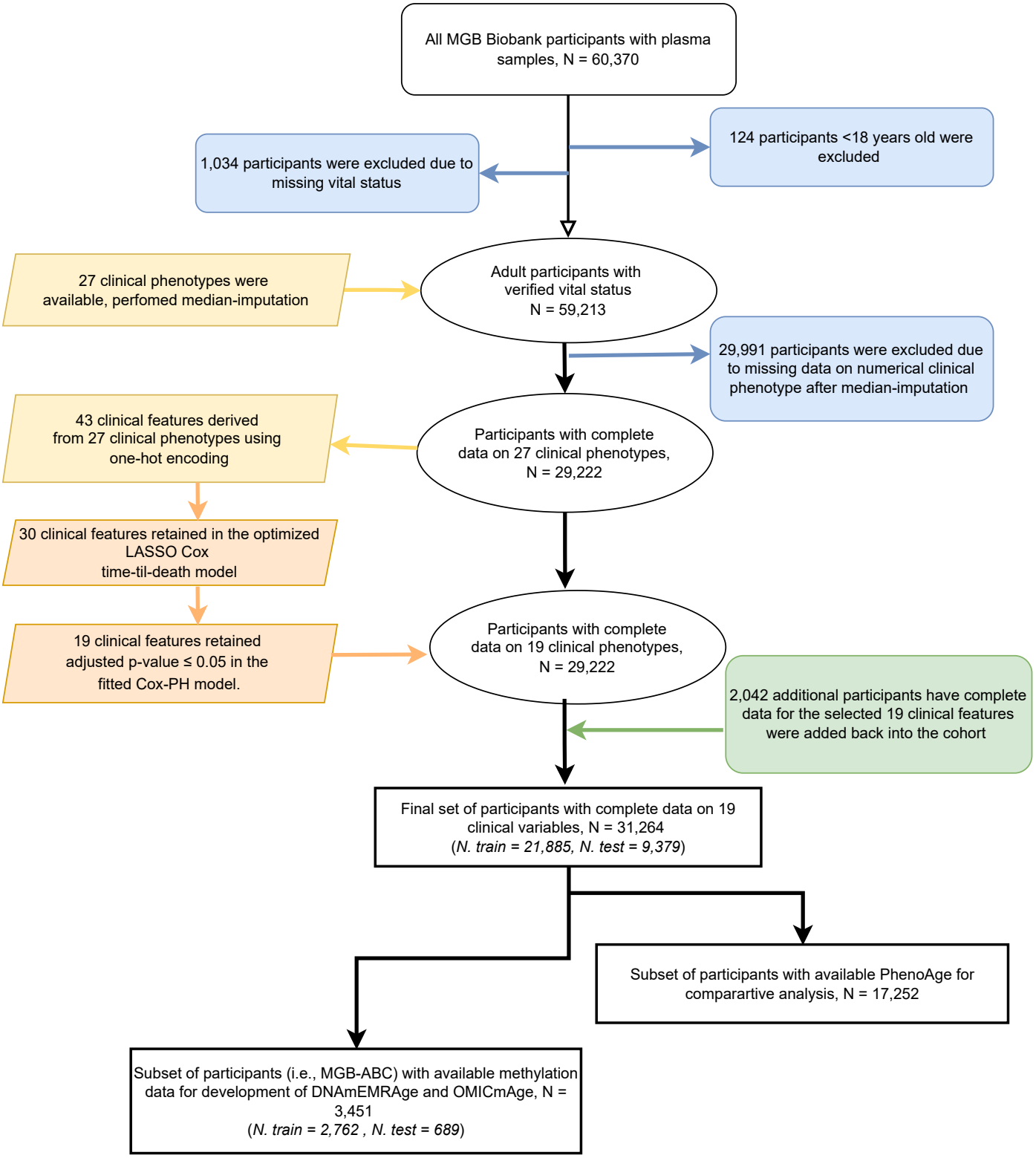
Peer review information *Nature Aging* thanks Daniel Belsky and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

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

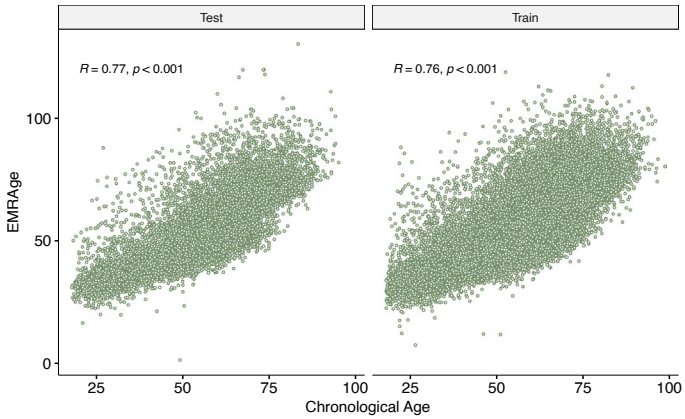
[1]Channing Division of Network Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [2]Department of Population Health Sciences, Duke University, Durham, NC, USA. [3]TruDiagnostic, Lexington, KY, USA. [4]Program in Computational Biology and Bioinformatics, Yale University, New Haven, CT, USA. [5]Department of Psychiatry, Yale University School of Medicine, New Haven, CT, USA. [6]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [7]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, MA, USA. [8]Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [9]Department of Health Informatics, Rutgers School of Health Professions, Rutgers Biomedical and Health Sciences, Newark, NJ, USA. [10]Division of Genetics, Department of Medicine, Brigham and Women's HospitalHarvard Medical School, Boston, MA, USA. [11]Department of Genetics, School of Medicine, Stanford University, Stanford, CA, USA. [12]Division of Endocrinology and Division of Genetics, Department of Medicine, BWH, Boston, USA. [13]Strathclyde Institute of Pharmacy and Biomedical Sciences, University of Strathclyde, Glasgow, UK. [14]Strathclyde Centre for Molecular Bioscience, University of Strathclyde, Glasgow, UK. [15]Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, MA, USA. [16]Unit of Integrative Metabolomics, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden. [17]Department of Respiratory Medicine and Allergy, Karolinska University Hospital, Stockholm, Sweden. [18]Division of Preclinical Innovation, National Center for Advancing Translational Science, National Institutes of Health, Rockville, MD, USA. [19]Department of Pathology, Yale University School of Medicine, New Haven, CT, USA. [20]These authors contributed equally: Qingwen Chen, Varun B. Dwaraka. ✉e-mail: rejas@channing.harvard.edu

https://doi.org/10.1038/s43587-026-01073-7



Extended Data Fig. 1 | Flowchart for inclusion of participants from the Massachusetts General Brigham (MGB) Biobank for the development of *EMRAge, DNAmEMRAge and OMICmAge*. Among 60,370 MGB Biobank participants with plasma samples, 59,213 adults with verified vital status were retained after excluding individuals under 18 years and those with missing survival data. After exclusion of participants with missing non-imputable clinical variables, 29,222 individuals with complete data on 27 clinical phenotypes remained. These were encoded into 43 features, of which 19 significantly associated with mortality were selected to construct EMRAge, yielding a final cohort of 31,264 participants (training $N = 21,885$; test $N = 9,379$). A subset with available DNA methylation and proteomic data ($N = 3,451$) was used to develop DNAmEMRAge and OMICmAge (training $N = 2,762$; test $N = 689$). An independent subset with available PhenoAge ($N = 17,252$) was used for comparative analyses.

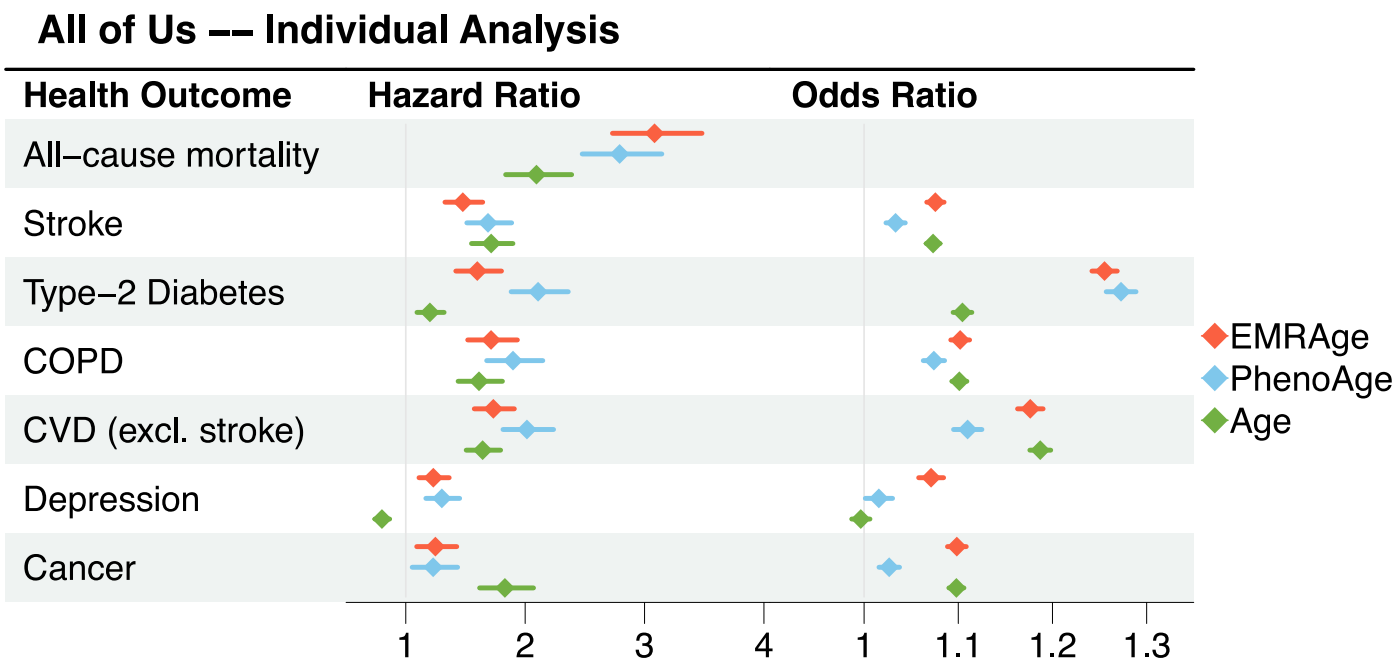
https://doi.org/10.1038/s43587-026-01073-7



Extended Data Fig. 2 | Pearson's correlation between *EMRAge* and chronological age in the MGB test and train sets. The Pearson correlation coefficient (ρ) was tested for significance using a two-sided t-test. The resulting P-values in both the MGB test and train sets were highly significant (P < 2.22E − 308).

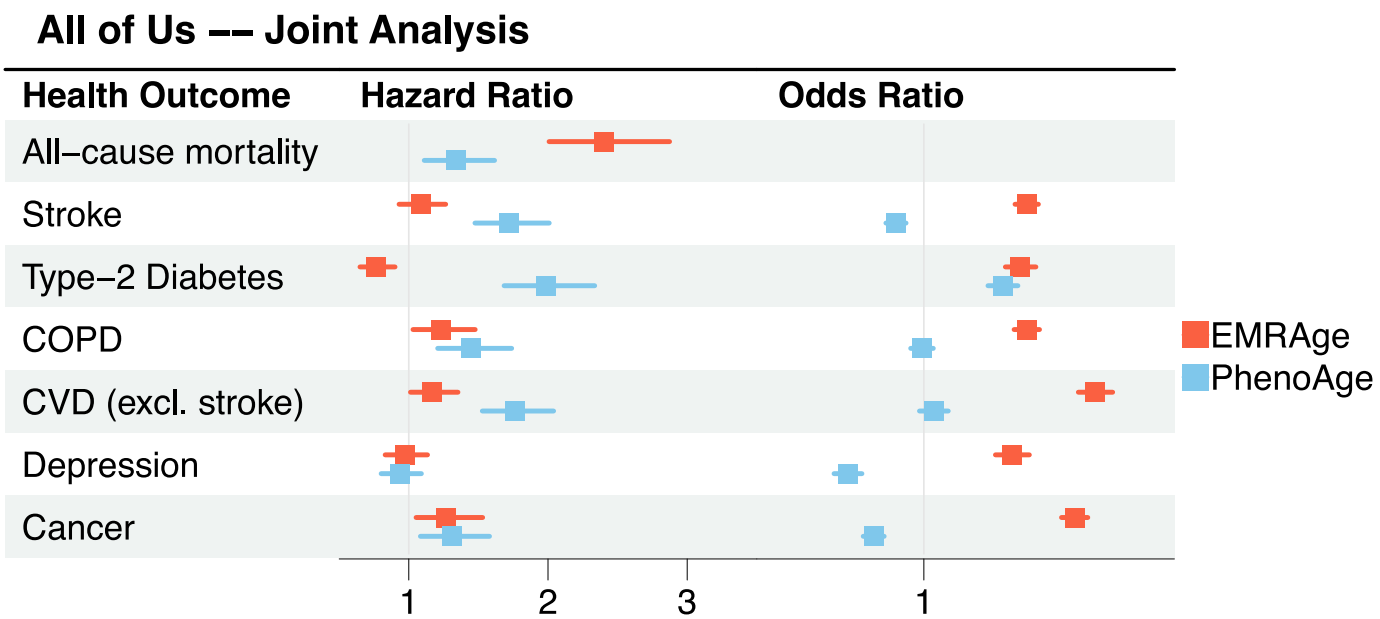
https://doi.org/10.1038/s43587-026-01073-7

3a)



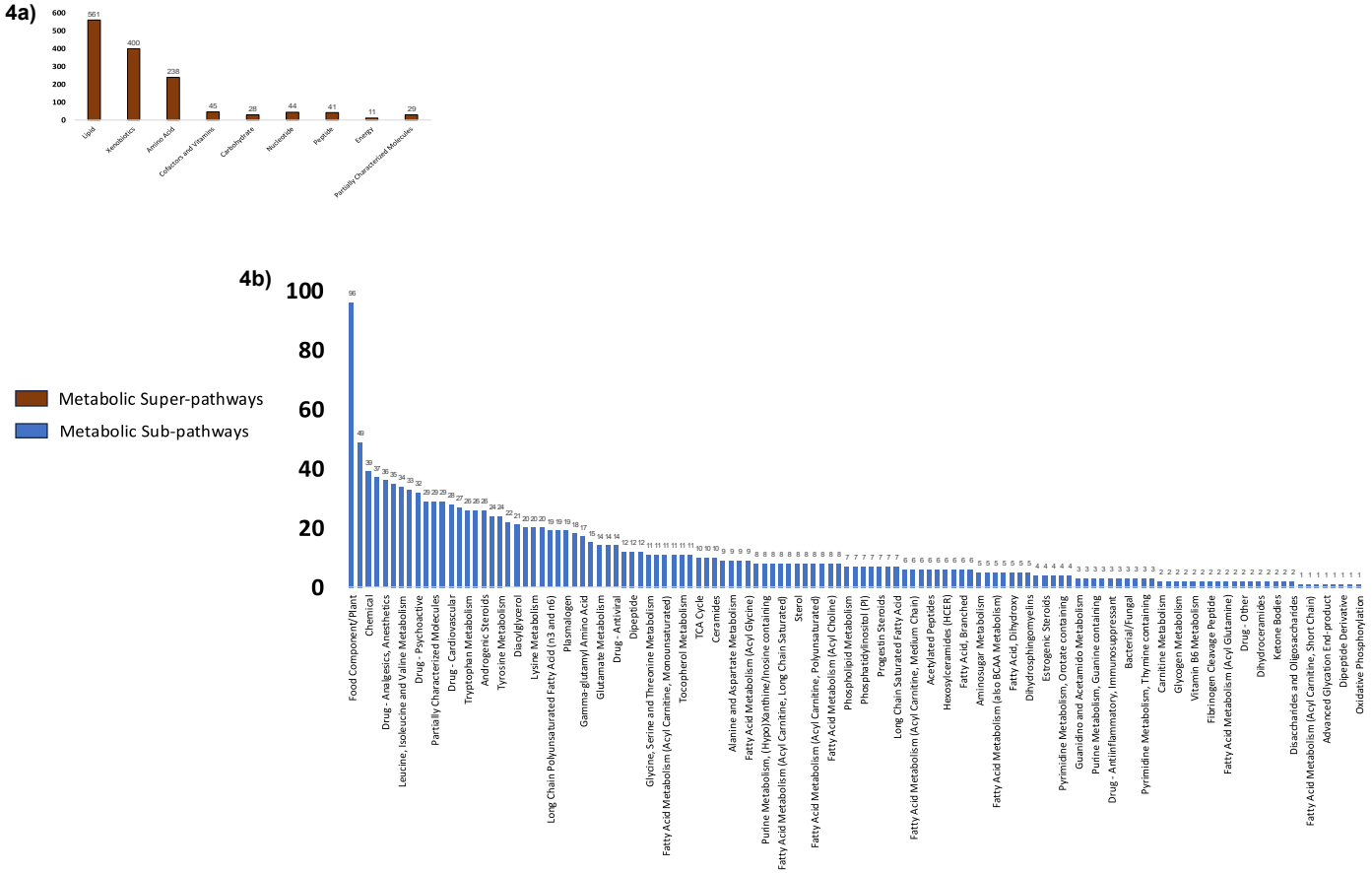
Adjusted for Age (in model for EMRAge/PhenoAge), Sex, Race, BMI, Smoking habit & Alcohol drinking status

3b)



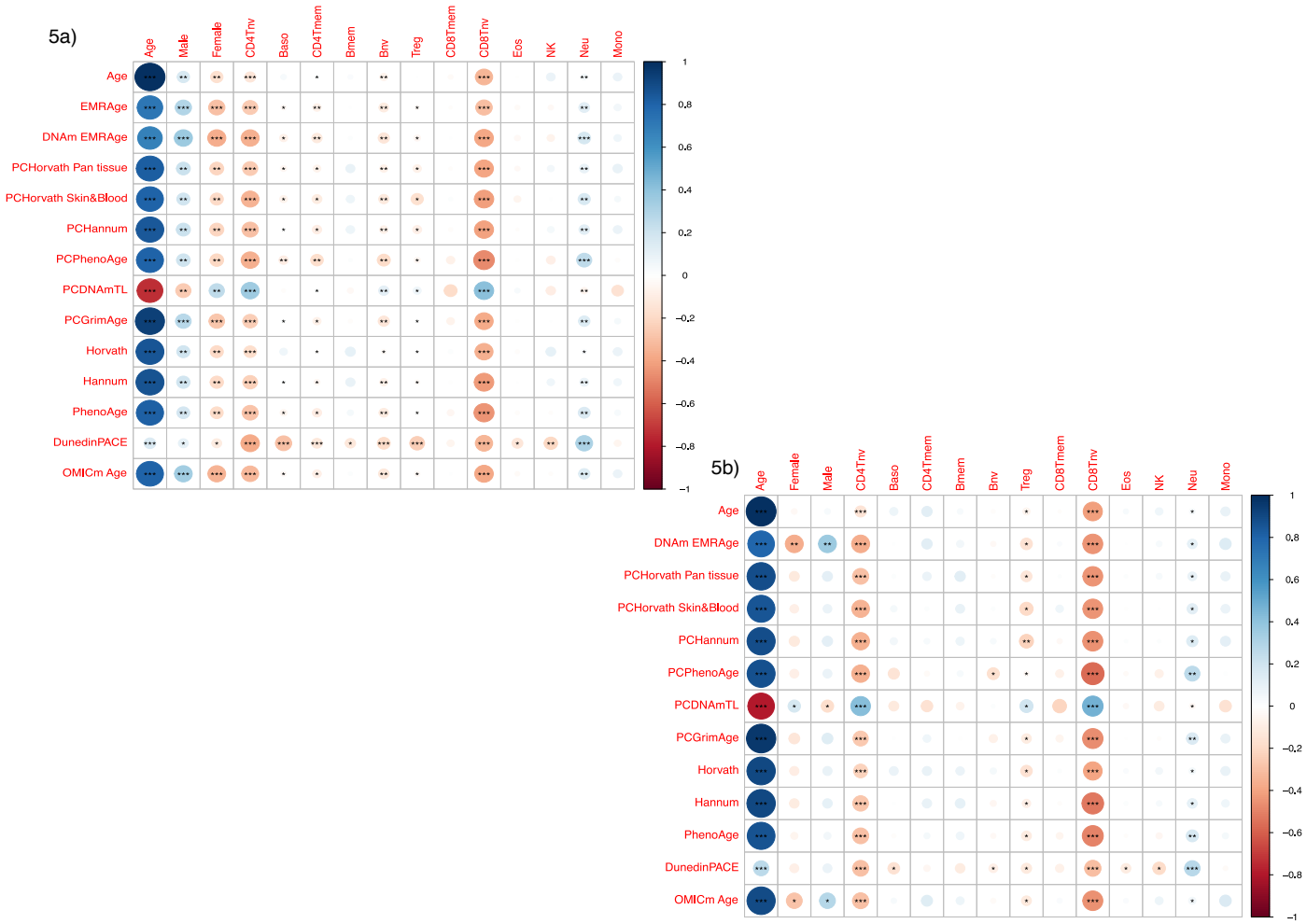
Extended Data Fig. 3 | Comparison of EMRAge and PhenoAge in association with aging-related health outcomes in the All of Us cohort. The analysis compares associations from individual (**a**) and joint (**b**) perspectives. Error bars show hazard/odds ratios and 95% confidence intervals per one-standard deviation change from the estimated mean value. For individual analysis, each aging metric (EMRAge, PhenoAge, or chronological age) was tested separately, adjusting for sex, race, BMI, smoking status, and alcohol consumption. Chronological age was also adjusted for when testing EMRAge or PhenoAge. For joint analysis, both EMRAge and PhenoAge were included as predictors in a single model, without additional adjustment. The number of cases (n) and sample sizes (N) for each phenotype are as follows: **Incident Cases**: all-cause mortality, N = 10,769 (n = 378); stroke, N = 9,285 (n = 548); type-2 diabetes, N = 6,946 (n = 503); COPD, N = 8,878 (n = 402); Depression, N = 6,226 (n = 639); CVD (excl. stroke), N = 4,734 (n = 658); cancer, N = 8,942 (n = 380). **Prevalent Cases**: The total sample size for prevalent diseases is 10,769, with the following case numbers: stroke, n = 1,484; type-2 diabetes, n = 3,823; COPD, n = 1,891; Depression, n = 4,543; CVD (excl. stroke), n = 6,035; cancer, n = 1,827.

https://doi.org/10.1038/s43587-026-01073-7



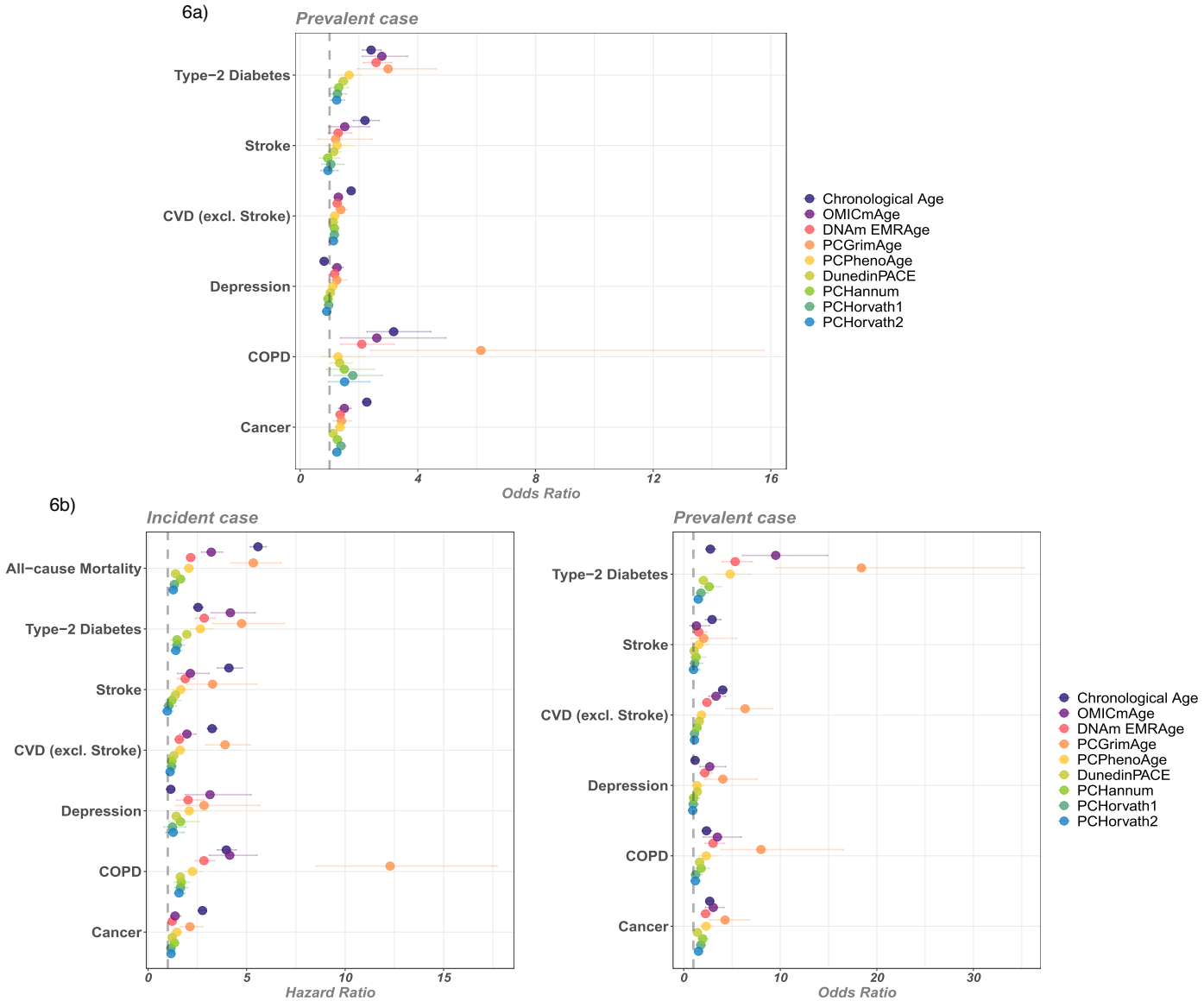
Extended Data Fig. 4 | Metabolite EBPs by super pathway and sub pathways. Distribution of super pathways (a) and sub pathways for all the metabolites (b). A total of 1,459 metabolites were categorized into nine super-pathways and sixty-three sub-pathways. The number of metabolites in each category is displayed above the respective bars.

https://doi.org/10.1038/s43587-026-01073-7



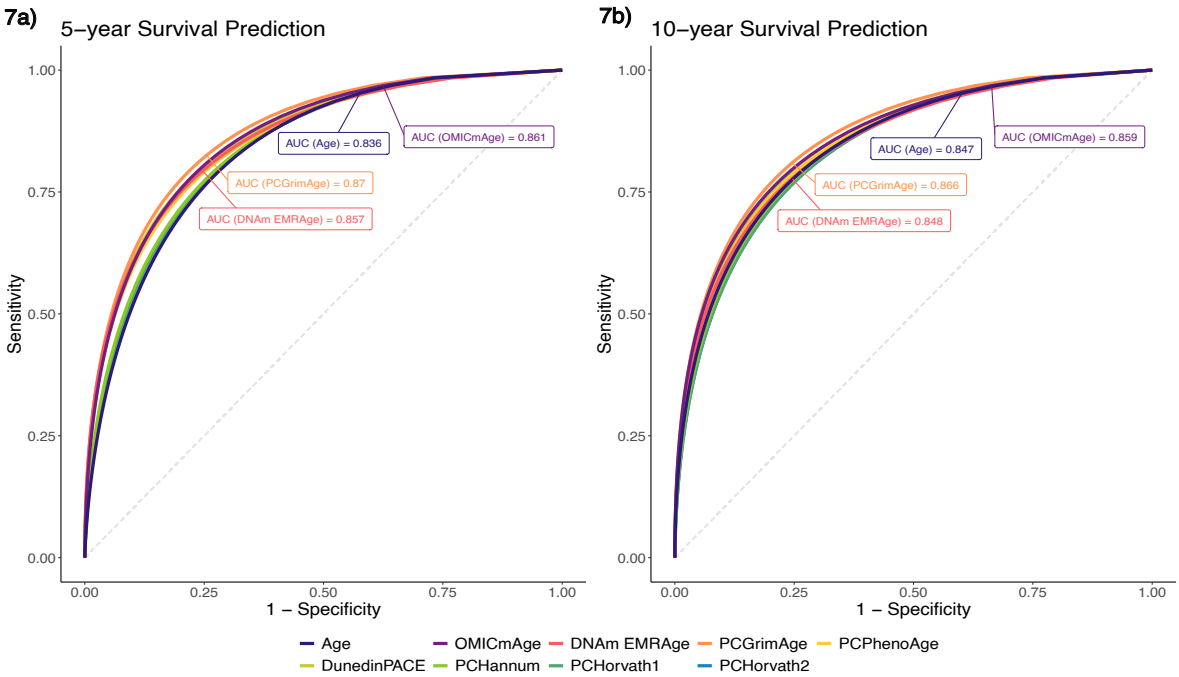
Extended Data Fig. 5 | Correlation plots between DNAm-based aging biomarkers and immune cell fractions in the MGB-ABC and TruDiagnostic Biobank cohorts. a) Discovery cohort - MGB-ABC biobank. **b**) Validation cohort - TruDiagnostic Biobank. Each column represents a covariate, including age, sex (male/female), and immune cell fractions. The size is adjusted to the magnitude of the correlation and the color to the direction (positive or negative). *: p-value < 0.05, **: p-value < 0.01, ***: p-value < 0.001. All Pearson correlation coefficients and associated P-values are included in the Source Data.

https://doi.org/10.1038/s43587-026-01073-7



Extended Data Fig. 6 | Horizontal error bar plot of odds/hazard ratios of each aging biomarker and chronological age to aging-related diseases and all-cause mortality in the two validation cohorts. a, Odds ratio of the TruDiagnostic Biobank. **b**, Hazard Ratio (left) and odds ratio (right) of the Generation Scotland. The ratios and 95% confidence intervals are based on a one-standard deviation change around the estimated mean values from the statistical models. Sample sizes (N) and the number of cases (n) for each phenotype are provided for the TruD (N_1), and GS (N_2) cohorts: **Incident case**: all-cause mortality, $N_2 = 18,672$ ($n_2 = 1,503$); type-2 diabetes, $N_2 = 18,488$ ($n_2 = 589$); stroke, $N_2 = 18,588$ ($n_2 = 314$); CVD (excl. stroke), $N_2 = 17,953$ ($n_2 = 1,026$); depression, $N_2 = 18,491$ ($n_2 = 172$); COPD, $N_2 = 18,535$ ($n_2 = 492$); cancer, $N_2 = 18,222$ ($n_2 = 1,203$). **Prevalent case**: type-2 diabetes, $N_1 = 14,213$ ($n_1 = 299$), $N_2 = 18,672$ ($n_2 = 184$); stroke, $N_1 = 14,213$ ($n_1 = 115$), $N_2 = 18,672$ ($n_2 = 84$); CVD (excl. stroke), $N_1 = 14,213$ ($n_1 = 5,046$), $N_2 = 18,672$ ($n_2 = 719$); depression, $N_1 = 14,213$ ($n_1 = 1,329$), $N_2 = 18,672$ ($n_2 = 181$); COPD, $N_1 = 14,213$ ($n_1 = 45$), $N_2 = 18,672$ ($n_2 = 137$); cancer, $N_1 = 14,213$ ($n_1 = 1,501$), $N_2 = 18,672$ ($n_2 = 450$).

https://doi.org/10.1038/s43587-026-01073-7



Extended Data Fig. 7 | ROC curves showing performance of survival prediction: 5-/10-year in the Generation Scotland cohort. ROC curves for 5-year (**a**) and 10-year (**b**) survival predictions in the GS cohort. Individual lines represent DNAm aging biomarkers compared against chronological age. 5-Year Survival AUCs: PCGrimAge (0.870), DunedinPACE (0.861), OMICmAge (0.861), DNAm EMRAge (0.857), PCPhenoAge (0.853), PCHannum (0.842), PCHorvath1 (0.839), PCHorvath2 (0.839), and chronological age (0.836). 10-Year Survival AUCs: PCGrimAge (0.866), OMICmAge (0.859), DunedinPACE (0.857), PCPhenoAge (0.852), DNAm EMRAge (0.848), chronological age (0.847), PCHannum (0.845), PCHorvath1 (0.841), and PCHorvath2 (0.841).

nature portfolio

Corresponding author(s): Jessica Lasky-SU

Last updated by author(s): Jan 6, 2026

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data was collected from private repositories, including those from clinical routines, epidemiological surveys, or molecular profiling of plasma samples.
Data analysis	For MGB cohorts: The omics data were quality controlled, processed and analyzed using R version 4.3.0 with packages including 'minfi', 'ENmix', 'dplyr', 'ggplot2', 'glmnet', 'survival', 'pROC' and 'stats', Platform: x86_64-pc-linux-gnu (64-bit), Running under: CentOS Linux 7 (Core). For TruDiagnostic cohort: Raw data was processed using the minfi (R package) pipeline. Low-quality samples were identified using the ENmix qcfilter() function. Probes with P-values<0.05 across all samples were identified and kept, with low-quality probesets removed. A combinatorial normalization processing using the minfi Funnorm procedure (available in the package minfi 1.44.0), followed by the RCP method (available in package ENmix 1.34.02). For GS cohort: DNA methylation data have been profiled using the Illumina EPICv1 array. Quality control details have been described previously60. Briefly, samples were assessed in four sets yielding data for 18,869 individuals after quality control (N-set1 = 5,087, N-set2 = 459, N-set3 = 4,450, N-set4 = 8,873). Here, after the removal of 11 individuals who subsequently withdrew consent, we had data for 18,858 volunteers. Secondary care linkage was available for 99% of GS volunteers (N-analysis = 18,672). OMICmAge and the other epigenetic biomarkers were estimated as described for the MGB-ABC cohort. We also included a Code Availability statament as follows: Code to calculate all metrics is accessible via TruDiagnostic's DNAm Analysis Software. You can request access to the software at https://www.trudiagnostic.com/softwarerequest, at no cost for research purposes. Access is contingent upon the requestor agreeing to a standard,

non-commercial software license agreement which confirms the intended use is strictly for academic, non-profit research and that the requestor will not attempt to reverse-engineer or commercially exploit the software. All R codes used to develop and validate EMRAge, DNAm EMRAge and OMICmAge are available at https://github.com/LaskySuLab/OMICmAge.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Research Patient Data Registry (RPDR) of the Mass General Brigham Biobank was queried using its online RPDR query tool. The details of the queried data terms are described and the references are cited in the text. Metabolomic profiling using non-targeted Liquid Chromatography Coupled Mass Spectroscopy (LCMS) platforms was performed for a subset of patients in MGB-ABC cohort by Metabolon Inc. (Durham, NC, USA). Proteomic profiling using Liquid Chromatography-tandem Mass Spectrometry (LC-MS/MS) platform was also performed for a subset of patients in MGB-ABC cohort by Seer Inc. (Redwood City, CA, USA). All participants provided written informed consent for biomedical research.

The TruDiagnostic cohort represent EPIC DNAm dataset from TruDiagnostic Inc. was collected between 2020 and 2023. The dataset represents whole blood samples collected from individuals who had provided blood as part of either a routine check by physicians or by acquiring a kit directly from TruDiagnostic Inc. All individuals have provided written informed consent to use the collected data for this project. Whole blood samples were collected and stored at −80°C prior to DNA processing, which was conducted at the TruDiagnostic Inc. lab facility (Lexington, KY, USA). Five hundred nanograms of DNA was extracted and bisulfite converted using the EZ DNA Methylation kit (Zymo Research) using the manufacturer's instruction. After bisulfite conversion, converted DNA were hybridized to the Illumina HumanMethylation EPIC Beadchip, stained, washed, and imaged with the Illumina iScan SQ instrument to obtain raw image intensities.

Generation Scotland (GS) is a Scottish, family-based cohort study with over 24,000 volunteers, aged 17-99, stemming from >5,500 families. The majority of volunteers provided blood samples at a baseline clinic between 2006 and 2011 in addition to completing health and lifestyle questionnaires and giving written informed consent for data linkage to their electronic health records. 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. This study was performed in accordance with the Helsinki declaration.

The All of Us (AoU) Research Program, an initiative launched by the NIH in 2018, represents a national collaborative effort to aggregate genetic, lifestyle, environmental, and electronic medical record (EMR) data from one million participants. All the participants provided a written informed consent form at the time of enrollment. Given the diverse data sources utilized by AoU, including EMR data from healthcare facilities, participant surveys, and self-reported measurements, the program implemented the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) to store and standardize this heterogeneous data, which were coded using the SNOMED CT dictionary. To facilitate the efficient retrieval of diagnosis records, we mapped the curated ICD-9 and ICD-10 codes to corresponding SNOMED CT codes. On February 4, 2025, AoU released the latest version of its Curated Data Repository (CDRv8), encompassing participant data up to a cutoff date of October 1, 2023. Within CDRv8, 389,379 adult participants had both EMR and lifestyle survey data available. Following that, we extracted demographic information, lifestyle factors, diagnosis records, and laboratory test results necessary for the calculation of EMRAge and PhenoAge. To address missing laboratory values and ensure a sufficient sample size for association analyses, we imputed missing values with the median of all measurements recorded within one year of each participant's enrollment date. This imputation process resulted in a final cohort of 10,769 participants with complete data from the All of Us Research Program.

We also included the Data Availability statement as follow:

The data used in this study were generated from two biobank cohorts housed at Mass General Brigham (MGB) and TruDiagnostic respectively. Requests for raw data, analyzed data and materials used in MGB and TruDiagnostic Biobank cohorts to generate the results in this study will be reviewed by the contact PIs (Jessica Lasky-Su or Varun Dwaraka). Due to participant consent restrictions, individual-level data linked to electronic health records cannot be made publicly available. However, we are committed to supporting scientific collaboration and transparency wherever possible. Researchers interested in non-commercial academic use of the data, algorithms, or derived biomarkers are encouraged to contact the corresponding author (Jessica Lasky-Su, rejas@channing.harvard.edu) or first author (Varun Dwaraka, varun@trudiagnostic.com). Upon review, a collaborative agreement or Data Use Agreement (DUA) can be established to enable:

- Access to de-identified omics data for research purposes
- Application of algorithms or generation of derived biomarker scores on external datasets
- Internal use of EHR-linked clinical data for validation, replication, or joint analyses with external collaborators

Requests will be assessed to determine whether they are subject to intellectual property or confidentiality obligations. Access to certain raw or analyzed data may be restricted if it relates to ongoing or future patent filings, or if release could compromise the commercialization strategy of MGB or TruDiagnostic. Access may also be limited for data containing proprietary algorithms or other commercially sensitive information. We welcome such requests and aim to respond within 30 days. Publicly available datasets referenced in this study include:

- All of Us Research Program: https://allofus.nih.gov
- Generation Scotland: https://genscot.ed.ac.uk
- SEER proteomics data for the MGB-ABC cohort: PRIDE accession PXD048709 (https://www.ebi.ac.uk/pride/archive/projects/PXD048709)

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	All but TruDianostic Biobank cohort represented a majority female group. Analyses were conducted with the population as a

Reporting on sex and gender	whole, and stratified between sexes. Sex and Age were adjusted for for all analyses and modeling. Detailed information is reported in Supplementary Table S1.
Reporting on race, ethnicity, or other socially relevant groupings	In the MGB Biobank cohort, most of patients self-reported as whites (84.0%), along with 5.8% reporting as African American, 2.3% as Asian, 5.5% reporting as other, and a small percentage (2.4%) of patients didn't report their race or ethnicity. In addition, the MGB ABC cohort represents a similar composition of race as MGB Biobank cohort. Among participants in TruDiagnostic cohort, majority of patients self-reported as white (76.7%), along with 5.9% reporting as Asian, 1.5% reporting as African American, and 15.9% reporting as other. In All of Us cohort, more than half of patients self-reported as white (65%), along with 10% reporting as African American, 1.6% reporting as Asian and 8.6% reporting as other. GS cohort includes the most White participants (96.4%). Detailed race characteristics are reported in Supplementary Table S1.
Population characteristics	All detailed Population characteristics are reported in Supplementary Table S1.
Recruitment	For MGB Biobank cohort, we identified all adult participants with complete phenotype information and available plasma samples as of July 28th, 2022. Written informed consent was obtained from all participants upon enrollment in the biobank. The MGB ABC cohort, a subset of the MGB Biobank cohort, includes patients whose samples were profiled for methylomics, metabolomics, or proteomics. All participants in both cohorts were recruited through Mass General Brigham healthcare centers. Therefore, both cohorts were more sick than a cohort not ascertained through clinical settings; however, our findings were validated in a distinct cohort comprising participants who are more proactive in self-care, which mitigates the potential selection bias due to higher illness severity observed in the MGB cohorts. The TruDiagnostic cohort represent EPIC DNAm dataset from TruDiagnostic Inc. was collected between 2020 and 2023. The dataset represents whole blood samples collected from individuals who had provided blood as part of either a routine check by physicians or by acquiring a kit directly from TruDiagnostic Inc. All individuals have provided written informed consent to use the collected data for this project. On February 4, 2025, AoU released the latest version of its Curated Data Repository (CDRv8), encompassing participant data up to a cutoff date of October 1, 2023. Within CDRv8, 389,379 adult participants had both EMR and lifestyle survey data available. Following that, we extracted demographic information, lifestyle factors, diagnosis records, and laboratory test results necessary for the calculation of EMRAge and PhenoAge. To address missing laboratory values and ensure a sufficient sample size for association analyses, we imputed missing values with the median of all measurements recorded within one year of each participant's enrollment date. This imputation process resulted in a final cohort of 10,769 participants with complete data from the All of Us Research Program. GS cohort includes all the volunteers providing blood samples at a baseline clinic between 2006 and 2011 in addition to completing health and lifestyle questionnaires and giving written informed consent for data linkage to their electronic health records. Then only participants with available Infinium Methylation EPIC BeadChip data were retained for analysis.
Ethics oversight	The study was approved by the following IRB institutions: Institute of Regenerative and Cellular medicine and Brigham and Women's Hospital

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal statistical power or sample size calculations were performed before data collection; instead, sample sizes were determined by practical considerations like participant availability and resource constraints. The MGB Biobank cohort recruited 31,264 participants with available plasma samples and complete clinical information by July 28, 2022, a large sample size considered sufficient due to its representation of the MGB system's patient population. The MGB-ABC cohort consists of 3,451 participants, balanced by age, sex, and BMI to create a representative aging population. The TruDiagnostic cohort includes 14,213 individuals who participated in epigenetic testing, representing a distinct, more health-conscious population recruited under healthcare recommendations. The All of Us cohort has 10,769 participants with complete clinical data for EMRAge and PhenoAge calculation. Finally, the GS cohorts includes 18,672 samples, limited to those with available Illumina EPICv1 data. Together, these sample sizes ensure representation across key demographic and experimental conditions, enhancing the generalizability of the findings within the scope of the study.
Data exclusions	To identify eligible participants for MGB Biobank cohort, we excluded 124 subjects because they're younger than 18 years old at the date of sample collection. We also excluded 1034 subjects due to missing information on their vital status as of 07/28/2022. The other 28329 subjects were also excluded due to missing phenotype information.
Replication	EMRAge was successfully replicated and its association with all aging-related diseases were validated in All of Us cohort. DNAmEMRAge and OMICmAge were both successfully replicated in GS and TruDiagnostic Biobank cohorts. The associations between DNAmEMRAge or

	OMICmAge and all-cause mortality, type-2 diabetes, CVD and COPD were validated in both cohorts. The cross-sectional association between DNAmEMRAge or OMICmAge and cancer were validated in both cohorts. The cross-sectional association between DNAmEMRAge or OMICmAge and depression was validated in the GS cohort.
Randomization	Subjects were chosen based on the availability of molecular data and phenotype information, therefore randomization was not applicable for the cohorts used in these study.
Blinding	Blinding is not applicable in this study due to its observational nature

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Plants

Seed stocks	Not applicable for this cohort.
Novel plant genotypes	Not applicable for this cohort.
Authentication	Not applicable for this cohort.