Title

A Metabolomic Aging Clock Using Human Cerebrospinal Fluid.

Journal

The Journals of Gerontology Series A, 77(4)

ISSN

1079-5006

Authors

Hwangbo, Nathan
Zhang, Xinyu
Raftery, Daniel
et al.

Publication Date

2022-04-01

DOI

10.1093/gerona/glab212

Peer reviewed

THE GERONTOLOGICAL SOCIETY OF AMERICA®

Journals of Gerontology: Biological Sciences
cite as: *J Gerontol A Biol Sci Med Sci*, 2022, Vol. 77, No. 4, 744–754
https://doi.org/10.1093/gerona/glab212
Advance Access publication August 12, 2021

OXFORD

Original Article

A Metabolomic Aging Clock Using Human Cerebrospinal Fluid

Nathan Hwangbo, BS,[1] Xinyu Zhang, PhD,[2] Daniel Raftery, PhD,[2,] Haiwei Gu, PhD,[2,] Shu-Ching Hu, MD, PhD,[3,4] Thomas J. Montine, MD, PhD,[5] Joseph F. Quinn, MD,[6,7] Kathryn A. Chung, MD,[6,7] Amie L. Hiller, MD,[6,7] Dongfang Wang, PhD,[2] Qiang Fei, PhD,[2] Lisa Bettcher, BS,[2] Cyrus P. Zabetian, MD, MS,[3,4] Elaine Peskind, MD,[3,8] Gail Li, MD, PhD,[3,8] Daniel E. L. Promislow, PhD,[9,10,] and Alexander Franks, PhD[1,*,]

[1]Department of Statistics and Applied Probability, University of California, Santa Barbara, USA. [2]Northwest Metabolomics Research Center, Department of Anesthesiology and Pain Medicine, University of Washington School of Medicine, Seattle, USA. [3]Veterans Affairs Puget Sound Health Care System, Seattle, Washington, USA. [4]Department of Neurology, University of Washington School of Medicine, Seattle, USA. [5]Department of Pathology, Stanford University School of Medicine, Palo Alto, California, USA. [6]Portland Veterans Affairs Medical Center, Oregon, USA. [7]Department of Neurology, Oregon Health and Science University, Portland, USA. [8]Department of Psychiatry and Behavioral Sciences, University of Washington School of Medicine, Seattle, USA. [9]Department of Biology, University of Washington, Seattle, USA. [10]Department of Laboratory Medicine and Pathology, University of Washington School of Medicine, Seattle, USA.

*Address correspondence to: Alexander Franks, PhD, Department of Statistics and Applied Probability, University of California, Santa Barbara, UCSB Statistics Department, 5607A South Hall, Santa Barbara, CA 93106, USA. E-mail: afranks@pstat.ucsb.edu

Received: March 30, 2021; Editorial Decision Date: July 13, 2021

Decision Editor: David Le Couteur, MBBS, FRACP, PhD

Abstract

Quantifying the physiology of aging is essential for improving our understanding of age-related disease and the heterogeneity of healthy aging. Recent studies have shown that, in regression models using "-omic" platforms to predict chronological age, residual variation in predicted age is correlated with health outcomes, and suggest that these "omic clocks" provide measures of biological age. This paper presents predictive models for age using metabolomic profiles of cerebrospinal fluid (CSF) from healthy human subjects and finds that metabolite and lipid data are generally able to predict chronological age within 10 years. We use these models to predict the age of a cohort of subjects with Alzheimer's and Parkinson's disease and find an increase in prediction error, potentially indicating that the relationship between the metabolome and chronological age differs with these diseases. However, evidence is not found to support the hypothesis that our models will consistently overpredict the age of these subjects. In our analysis of control subjects, we find the carnitine shuttle, sucrose, biopterin, vitamin E metabolism, tryptophan, and tyrosine to be the most associated with age. We showcase the potential usefulness of age prediction models in a small data set (n = 85) and discuss techniques for drift correction, missing data imputation, and regularized regression, which can be used to help mitigate the statistical challenges that commonly arise in this setting. To our knowledge, this work presents the first multivariate predictive metabolomic and lipidomic models for age using mass spectrometry analysis of CSF.

Keywords: Aging clock, Biomarker, Cerebrospinal fluid, Metabolomics

Risk of age-related disease varies among individuals and is shaped by genetic factors, environmental factors, and the interaction between the 2 (1). As with most complex traits, in attempts to map specific genetic variants that are associated with phenotypic variation, researchers have identified genes that are significantly correlated with age-related disease, but which explain only a small fraction of the overall variation (2,3), contributing to the so-called "missing heritability" problem (4). In order to address the large functional distance between genotype and phenotype, researchers have turned to "endophenotypes"—transcriptome, epigenome, metabolome,

lipidome, and microbiome—as a way to bridge the gap between genotype and phenotype, and characterize the physiological processes of aging (5).

In our own recent work, we have focused on the metabolome, which measures the structural and functional building blocks of an organism, as a powerful link between genotype and phenotype in studies of aging and age-related traits (6,7). During this same period, there has been considerable interest in the power of the epigenome to explain variation in patterns of aging within human populations. These prior studies have established predictive models for age using epigenetic data (8,9), and these "epigenetic clocks" have been shown to be useful biomarkers for risk factors of mortality (10). Motivated by these findings, many in the field of aging research have suggested that the epigenome can be used as a biomarker of underlying physiological age and so considered as a "biological clock."

Here, we focus on the degree to which variation in the metabolome, specifically as measured in cerebrospinal fluid (CSF), might function as a "metabolomic clock." Researchers have found that variation in the metabolome, which measures small molecules (<2 000 Da) circulating in an organism, can account for variation in diverse traits, including all-cause mortality (11,12). Furthermore, individual and small sets of metabolite concentrations have been found to be associated with age (13,14). Rather than focus on the association between a single metabolite and age, we use complete metabolomic profiles to create predictive models for age. Because the metabolome is the endophenotype furthest downstream in the genotype–phenotype path, a predictive model for age using the metabolome might provide unique insight into the physiological mechanisms of aging.

Metabolomic clocks using urine, serum, and plasma samples have been found to be predictive of chronological age (15–18). In the present study, we construct predictive models for age analyzing both targeted and untargeted metabolomic and lipidomic profiles of CSF. CSF serves to cushion the brain, as well as transport biological substances, and is the fluid in closest proximity to the central nervous system, making it valuable for analysis, particularly with respect to age-related neurological disease (19). There is some evidence that the relationship between metabolites in the brain and age is distinct from other organs, suggesting that a predictive model using CSF might provide new insight into aging in the brain (20,21).

The invasive nature of collecting CSF typically means that large sample sizes are difficult to obtain. To our knowledge, this is the first attempt at creating a predictive model for age using CSF data. This study is also notable for the comprehensive set of metabolomic profiles analyzed here, including targeted metabolomics, global metabolomics, and lipidomic profiles. Previous efforts to create biological clocks from high-dimensional data (8,15) have relied on a class of statistical models known as *Elastic Net* regularized linear regression. Such models are commonly used to form these age prediction models because of their ability to handle the case when the number of features (eg, methylation sites, metabolites) greatly exceeds the number of samples, and to perform feature selection in the process (16,22,23). We discuss techniques and challenges involved in fitting these models in small sample, high-dimensional profiles, including methods for missing data imputation and cross-validation, as well as methods to mitigate bias that can occur when fitting these high-dimensional regression models. We also address broader statistical challenges in these -omic based "biological clock" analyses, such as accounting for batch effects and signal intensity drift over time. We carry out pathway analysis and Metabolite Set Enrichment Analysis (MSEA) using univariate relationships between metabolites and age and find that the carnitine shuttle, sucrose, biopterin, vitamin E metabolism, tryptophan, and tyrosine are associated with age. Finally, in addition to analyzing healthy individuals, we also use our models to assess any evidence of accelerated aging in the metabolome profiles of patients with Alzheimer's disease (AD) and Parkinson's disease (PD).

Materials and Methods

Biological Samples

One hundred ninety-eight CSF samples, collected from 85 controls, 57 AD patients, and 56 PD patients, were available for analysis. The 85 healthy subjects range from 20 to 86 years of age with a median of 56, while the AD and PD subjects' range from 35 to 88 years of age with a median of 67. We also record each subject's sex at birth. Data on apolipoprotein-ε4 allele (*APOE* ε4) genotype (all participants) and glucocerebrocidase (*GBA*) carrier status (pathogenic mutations and the E326K polymorphism for the PD group only) were generated in previous work by our group and made available for this study (24,25).

For the control and AD subjects, CSF samples were provided from the Veterans Affairs (VA) Northwest Mental Illness Research, Education, and Clinical Center Sample and Data Repository. Participants underwent neurological examination and a detailed neuropsychological assessment. Participants were determined to be cognitively normal controls or AD patients by expert clinical diagnosis confirmed by neuropsychological testing and Clinical Dementia Rating scale. All control participants' samples were banked from a cross-sectional study of *APOE* genotype on CSF level of Abeta42 across the cognitively normal adult life span. Plasma, serum, and CSF were banked for use in future studies related to aging and neurodegenerative disorders. All control subjects were medically healthy, cognitively normal volunteers recruited from the community, with no neurologic or psychiatric disorders or complaints. Control participants had a Clinical Dementia Rating scale score of 0, Mini-Mental State Examination scores between 26 and 30, and paragraph recall scores not less than 1 *SD* below age- and education-matched norms. AD participants met National Institute of Neurological and Communicative Diseases and Stroke–Alzheimer's Disease and Related Disorders Association's criteria for probable AD.

CSF from PD subjects was obtained from participants enrolled in the Pacific Udall Center at the VA Puget Sound Health Care System and the VA Portland Health Care System. All PD subjects underwent a neurological examination and a detailed neuropsychological assessment. The resulting data were reviewed at a joint consensus conference to ensure diagnostic consistency between the 2 sites as previously described (26,27). Only participants who met UK Parkinson's Disease Society Brain Bank clinical diagnostic criteria for PD were included in this study (28).

At all sites, CSF was collected in the fasting state in the morning. CSF was collected in the lateral decubitus position using the Sprotte 24g atraumatic spinal needle. It was collected by negative pressure into sterile polypropylene syringes and aliquoted into 0.5 mL aliquots in polypropylene cryotubes and frozen immediately on dry ice at the bedside. All samples were stored at –80 °C prior to assay.

The study procedures were approved by the institutional review boards of the University of Washington, VA Puget Sound Health Care System, and VA Portland Health Care System. All participants, or their legally authorized representative for those with impaired decisional capacity, provided written informed consent.

Metabolomics

Samples were aliquoted into 4 different subsamples and, subsequently, prepared for 4 distinct metabolomic profiles, with 3 aqueous metabolite profiles, including (i) targeted metabolomics, (ii) global metabolomics, (iii) globally optimized targeted mass spectrometry (GOT-MS), and (iv) lipidomics.

Targeted metabolomics

Targeted metabolomics analysis was carried out using an liquid chromatography–tandem mass spectrometry (LC–MS/MS) platform targeting 203 standard metabolites from more than 25 metabolic pathways (eg, glycolysis, tricyclic acid cycle, amino acid metabolism, glutathione, etc.). LC–MS/MS experiments were performed on a Waters Acquity I-Class UPLC TQS-micro MS system. Each sample was injected twice, 2 and 10 µL, for analysis using positive and negative ionization modes, respectively. Both chromatographic separations were performed in hydrophilic interaction chromatography mode. The flow rate was 0.3 mL/min, autosampler temperature was kept at 4 °C, and the column compartment was set at 40° C. The mobile phase was composed of Solvents A (5 mM ammonium acetate in H_2O + 0.5% acetic acid + 0.5% acetonitrile) and B (acetonitrile + 0.5% acetic acid + 0.5% water). The LC gradient conditions were the same for both positive and negative ionization modes. After an initial 1.5-minute isocratic elution of 10% A, the percentage of Solvent A was increased linearly to 65% at time (t) = 9 minutes, then remained the same for 5 minutes (t = 14 minutes), and then reduced to 10% at t = 15 minutes to prepare for the next injection. After chromatographic separation, MS ionization and data acquisition were performed using an electrospray ionization (ESI) source. A pooled study sample was used as the quality control and run once for every 10 study samples.

Global metabolomics

Global MS-based lipidomics was performed using an Agilent 1200 LC system coupled to an Agilent 6520 quadrupole-time-of-flight mass spectrometer (Q-TOF/MS). CSF samples (200 µL) were prepared by dissolving in 1 000 µL methanol, vortexed, incubated at –20 °C, and centrifuged at 18 000 g for 20 minutes. Then 750 µL was collected, dried and then reconstituted in a 100 µL solution of 40% H_2O 60% acetonitrile. Five microliters of each prepared sample were analyzed by positive ESI and 10 µL was analyzed by negative ESI. Samples were separated using a Waters XBridge BEH Amide column (15 cm × 2.1 mm, 2.5 µm), which was heated to 40 °C. The mobile phase was 10 mM NH_4HCO_3 in 100% H_2O (Solvent A) and 100 % acetonitrile (Solvent B) and its gradient was 95%–10% B from 0 to 5 minutes, 10% B from 5 to 40 minutes, 10%–100% B from 40 to 45 minutes, and 100% from 40 to 70 minutes. MS parameters were performed according to previously described methods (29). The mass accuracy of our LC–MS system is generally better than 5 ppm, the Q-TOF/MS spectrometer was calibrated prior to each batch run, and a mass accuracy of <1 ppm was often achieved using the standard tuning mixture (G1969-85000, Agilent Technologies). The m/z scan range was 100–2 000, and the acquisition rate was 1.0 spectra/s. MS data were processed using Agilent Mass Hunter and Mass Profiler Professional.

Globally optimized targeted mass spectrometry

GOT-MS is a technique which sits between targeted and untargeted metabolomic profiling (30,31). The GOT-MS method used here was modeled after Zhong et al. (32), and is detailed in Supplementary Methods.

Lipidomics

Lipids were extracted from the samples (200 µL) using dichloromethane/methanol after the addition of 54 isotope-labeled internal standards across 13 lipid classes. The extracts were concentrated under nitrogen and reconstituted in 100 µL solution consisting of 10 mM ammonium acetate in dichloromethane:methanol (50:50). Lipids were analyzed using the Sciex Lipidyzer platform consisting of a Shimadzu Nexera X2 LC-30AD pumps, a Shimadzu Nexera X2 SIL-30AC autosampler, and an AB Sciex QTRAP 5500 MS/MS system equipped with SelexION for differential mobility spectrometry (DMS) according to the methods we developed previously (33). Multiple reaction monitoring was used to target and quantify over 1 000 lipids in positive and negative ionization modes with and without DMS. Data were acquired and processed using Sciex Analyst 1.6.3 and Lipidomics Workflow Manager 1.0.5.0.

Data Analysis

Data preprocessing

Samples were prepared for assay in 7 batches of 28 or 29 samples each. To mitigate batch effects, we explicitly balanced age, sex, phenotype (controls, Alzheimer's, Parkinson's), and APOE status across batches. To this end, we used a variant of the "finite selection model" proposed in Morris (34). For each of the 7 batches, we iteratively selected the subject that was the most different from the subjects already assigned to that batch, with respect to the aforementioned covariates. This process balances the samples across the batches, which is important for correcting drift and maximizing power to detect differences. Covariate balance across batch is shown in Supplementary Figure 1.

Prior to inference, we corrected for systematic drift in the mass spectrometer inferred intensities over time. The observed drifts were metabolite-specific and thus separate corrections were done for each metabolite. In Supplementary Figure 2, we show examples of metabolites for which the raw intensities showed significant changes in mean intensity over time, along with the data after correction. To account for the very general functional form of the intensity drift over time, we used extreme gradient boosting, a tree-based ensemble method that can recover noncontinuous functions. We use the xgboost package in R and cross-validation to select tuning parameters and control overfitting (35).

After drift correction, we centered and scaled each feature, and treated observations further than 3 median absolute deviations within each feature as missing values to be imputed. Features with more than 50% missingness were removed. A comparison of missingness across profiles is displayed in Figure 1.

Missing data imputation

We use the R package Amelia to estimate the missing values in our data set (36). This method creates multiple completed versions of the data set under the assumption that the data are well approximated by a multivariate normal distribution, and that the missing values are missing at random, which is to say, the differences between the missing and observed values can be fully explained using the observed variables in the data set. We created 5 imputed data sets of each profile and ran our analysis on each of these completed data sets. The variation across imputed data sets provides a measure of robustness to the imputation procedure. In Figure 2, the error bars

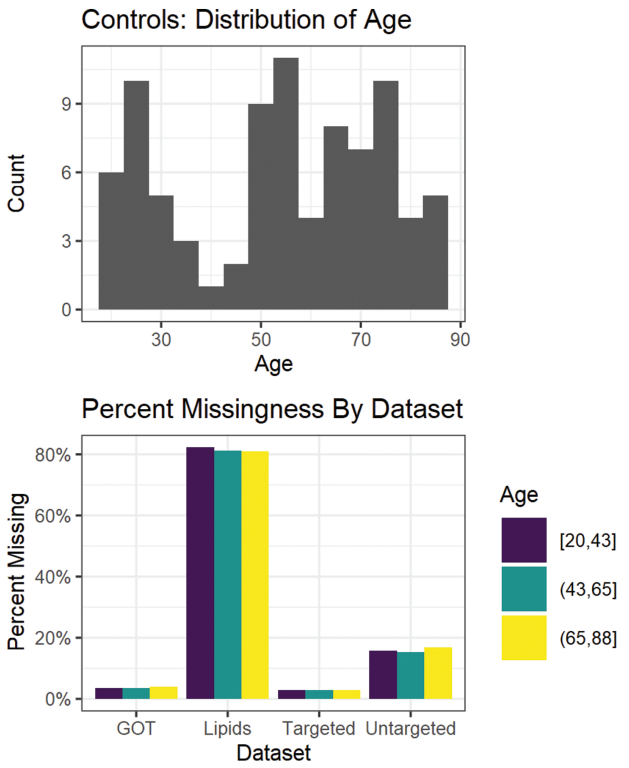

Figure 1. Distribution of age (top) as well as a comparison of missingness between profiles, split in 3 approximately equal-sized age groupings (bottom). GOT = globally optimized targeted. Full color version is available within the online issue.

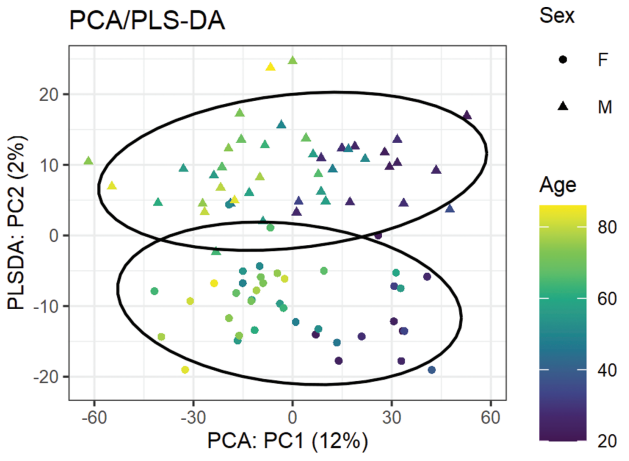

Figure 2. Separation of sex and age using the combined result of PCA and PLS-DA on the data set of combined profiles. The x-axis displays the first principal component from PCA, while the y-axis displays the first PLS component discriminating sex at birth on the space orthogonal to x-axis. For PLS-DA, the first 2 principal components are plotted, with confidence ellipses (assuming t distribution). In parentheses, the axis names contain the percent variation (of the data set) explained by that component. PCA = principal component analysis; PLS-DA = partial least squares discriminant analysis. Full color version is available within the online issue.

indicate the minimum, maximum, and average result among the 5 imputations.

Because this method requires that the number of subjects exceeds the number of metabolites, we assume that each metabolite follows a normal distribution, so that the underlying distribution of the transposed data is multivariate normal. In cases of large amounts of missingness, we add a ridge prior, shrinking the assumed covariances between metabolites. These modifications make it possible to complete the imputation at cheaper computational cost, but add additional bias to the imputed values.

Principal component analysis and partial least squares discriminant analysis

Principal component analysis (PCA) is a technique commonly used for visualizing high-dimensional data by projecting it onto a lower-dimensional space. This unsupervised procedure identifies dimensions that maximize the amount of preserved variability in the data. In addition to PCA, we can also make use of a supervised method known as partial least squares. Partial least squares maximizes the covariance between the data and a given variable. Partial least squares discriminant analysis (PLS-DA) is the special case where the provided variable is categorical. Both PCA and PLS-DA were implemented using the Nonlinear Iterative Partial Least Squares algorithm in the R package ropls, which allows for missing values (37). We combine these 2 techniques in the following way:

- Perform PCA on the combined profiles data set X to obtain a loading matrix W and score matrix T, which satisfies $X = TW^T$.
- Extract the first column of T (t_1) and the first row of W (w_1^T) and compute $\tilde{X} := X - t_1 w_1^T$.
- Run PLS-DA on $\tilde{X}$ with respect to sex and plot the first PLS component against the first principal component from PCA (t_1).

Regression modeling

Gaussian elastic net regression models extend the linear regression model to allow for the case where the number of features exceeds the number of observations, and can be implemented using the glmnet package in R (22). Two parameters need to be specified by the user to implement this model: α and λ. Following Horvath (8), we use α = .5, which performs variable selection but avoids randomly choosing one out of a set of correlated predictors (a problem with models fit with α = 1). Values of λ, the penalty parameter, are chosen in each model to minimize squared error in leave-one-out cross-validation (LOOCV).

The only metadata included in the model is the sex of each subject. Because of the putative sex-related differences in the metabolome (16,38), we assume sex at birth is a priori an important predictor and do not perform shrinkage for this variable.

To evaluate the performance of these models, we proceed in a leave-one-out fashion by iteratively omitting a single observation and training a model on the remaining data. For each observation, we proceed as follows:

1. Remove features missing from the left-out observation from the training set, following the reduced modeling methodology validated in Saar-Tsechansky and Provost (39).
2. Impute the remaining missing values from the training set to create 5 imputed versions of the data set.
3. Fit models on the 5 imputed training sets, using LOOCV to tune λ.
4. Predict the age of the left-out observation in each imputation.

Applying this process to every observation and comparing the predictions to chronological age yield the plots seen in Figure 2. To

obtain a snapshot of the features driving these predictions, a model is fit on all the control subjects. The coefficients for this full model in the targeted and lipidomic profiles are reported in Table 1. The full model on the untargeted profile is then used to predict the age of an out-of-sample cohort containing 57 AD and 56 PD subjects. These patients tended to be older and in a smaller age range than

Table 1. Names and Average Coefficient for Targeted Metabolites (left) and Lipids (right) Retained in Elastic Net Models Fit on the Full Data

Metabolite	Coefficient (positive)	Lipid	Coefficient (positive)
Xanthine	3.7	SM(18:1)	8.73
Kynurenine	3.36	SM(16:0)	5.74
Carnitine	2.91	TAG52:2-FA18:1	3.18
HIAA	2.88	DCER(24:1)	2.75
Cystine	2.11	FFA(24:0)	2.28
Glycylproline	1.14	SM(14:0)	2.04
Aspartic acid	1.03	PE(18:1/18:1)	0.88
Glycine	0.99	CE(22:5)	0.86
Acetylcarnitine	0.94	PC(16:0/16:0)	0.84
Serotonin	0.93	TAG52:2-FA16:0	0.37
Caffeine	0.87	FFA(18:3)	0.29
3α-Hydroxy-12 ketolithocholic acid	0.85	PC(18:1/20:4)	0.27
DOPA	0.82	PC(16:0/16:1)	0.24
Anthranilic acid	0.37	FFA(20:1)	0.19
Decanoylcarnitine	0.31	PC(16:0/14:0)	0.19
2-Deoxyguanosine	0.13	TAG50:3-FA16:0	0.16
Inosine	0.08	TAG52:3-FA18:1	0.09
1-Methyladenosine	0.04	PE(18:0/22:6)	0.03
Acetylglycine	0.02		
Amiloride	0.01		
Metabolite	Coefficient (negative)	Lipid	Coefficient (negative)
4-Aminobutyric acid	−2.63	HCER(24:1)	−4.4
Serine	−2.26	PE(O-18:0/22:4)	−3.01
Uridine	−2.01	TAG55:5-FA18:1	−1.76
Acetamide	−1.89	TAG46:0-FA16:0	−1.56
Citraconic acid	−1.39	TAG56:8-FA20:4	−1.46
Adenosine	−1.22	DAG(14:0/14:0)	−1.4
Phenylalanine	−1.17	PE(18:0/22:4)	−1.21
Asparagine	−1.03	PE(P-18:0/20:4)	−1.05
Fructose	−0.76	FFA(14:1)	−0.99
Uracil	−0.68	PE(18:0/18:1)	−0.92
Alpha-hydroxyisovaleric acid	−0.52	FFA(20:4)	−0.79
Alpha-ketoisovaleric acid	−0.5	LPC(18:1)	−0.74
4-Methylvaleric acid	−0.28	HCER(24:0)	−0.67
Arginine	−0.17	TAG52:6-FA18:1	−0.55
Glycocyamine	−0.16	CE(20:2)	−0.55
Tryptophan	−0.05	LPC(16:0)	−0.43
Homoserine	−0.03	TAG54:6-FA18:2	−0.42
Glycerol 3-phosphate	−0.03	TAG52:7-FA16:0	−0.35
		HCER(22:0)	−0.34
		FFA(20:5)	−0.26
		PC(16:0/20:1)	−0.24
		TAG48:1-FA18:1	−0.22
		TAG48:0-FA14:0	−0.21
		TAG51:0-FA18:0	−0.19
		PC(18:0/18:2)	−0.08
		PE(P-16:0/20:4)	−0.05
		PE(18:0/20:4)	−0.03
		CER(24:1)	−0.02
		TAG56:7-FA20:4	−0.01
		TAG48:0-FA16:0	−0.01
		PC(14:0/18:1)	−0.01

Notes: CE = cholesterol ester; CER = ceramides; DAG = diacylglycerol; DCER = dihydroceramides; DOPA = dopamine; FFA = free fatty acids; HCER = hexosylceramides; HIAA = hydroxy-indoleacetic acid; LCER = lactosylceramide; LPC = lysophosphatidylcholine; LPE = lysophosphatidylethanolamine ; PC = phosphatidylcholine; PE = phosphatidylethanolamine; SM = sphingomyelin; TAG = triacylglycerol. The table is sorted by magnitude and split into positive and negative coefficients. Sex at birth was not penalized and is, therefore, on a different scale from the coefficients in the table. With this in mind, being male led to a decreased age prediction of 0.90 years in the targeted profile and a decreased age prediction of 2.87 years in the lipid profile.

our healthy cohort. Thus, rather than comparing the predictions for the AD/PD group using error metrics for the full set of controls, we opt for the following procedure:

1. For each subject in the new cohort, find the control with the same sex and closest age. If there are ties, then pick the subject with the closest processing batch number.
2. Using the untargeted model fit on the full set of controls, compute a predicted age for each of the AD/PD subjects.
3. Compare results between the 2 sets.

Pathway Analysis

Mummichog is a tool used for pathway analysis of untargeted metabolomic data, accessible via a script using the Python programming language (40). To use this program, we input mass:charge ratio and retention time for each feature as well as t values and (Benjamini–Hochberg corrected) p values taken from a univariate linear model regressing age on each feature. The output is a comparison of the input with known metabolomic pathways, split by the features' mode.

For pathway analysis on the targeted profile, MSEA is a tool which finds associations between sets of metabolites by performing a hypergeometric test (41). MSEA is implemented using the MetaboAnalystR package, which takes as input the names of significant metabolites (using a false discovery rate [FDR] <.05 from a univariate linear regression), as well as the names of all the features in our data set to be used as a reference set (42).

Results

The samples were processed using untargeted and targeted metabolite profiling, as well as GOT-MS and lipid profiling. The untargeted, targeted, GOT-MS, and lipid profiling yielded 6 735, 108, 854, and 1 070 features, respectively. Supplementary Figure 3 displays a flowchart outlining the analysis performed on each of the 4 profiles.

Principal component analysis/PLS-DA

In a data set formed by combining all 4 profiles, PCA showed that the first 2 principal components explain 20% of the variance in our data set. Subjects with negative scores for the first principal component have a median age of 70, while subjects with a positive score have a median age of 28, indicating that the largest source of variation in the profiles is driven by subject age. We also apply PLS-DA (a supervised method) on the combined data set in order to simultaneously separate the profiles by sex. We find that these first 2 PLS components account for only 11% of the variation in our data. Following the procedure described in the "Principal component analysis and partial least squares discriminant analysis" section, Figure 3 shows the combined data projected onto the first principal component from PCA (PC1), along with the first PLS component discriminating sex on the orthogonal complement of PC1.

Missing Data

Each of the 4 metabolome profiling techniques yielded widely varying degrees of missing data, with 16%, 3%, 4%, and 81% of the concentration values missing from the untargeted, targeted, GOT-MS, and lipid profiles, respectively. We find that the youngest patients tended to have the largest amount of missing data across all the profiles,

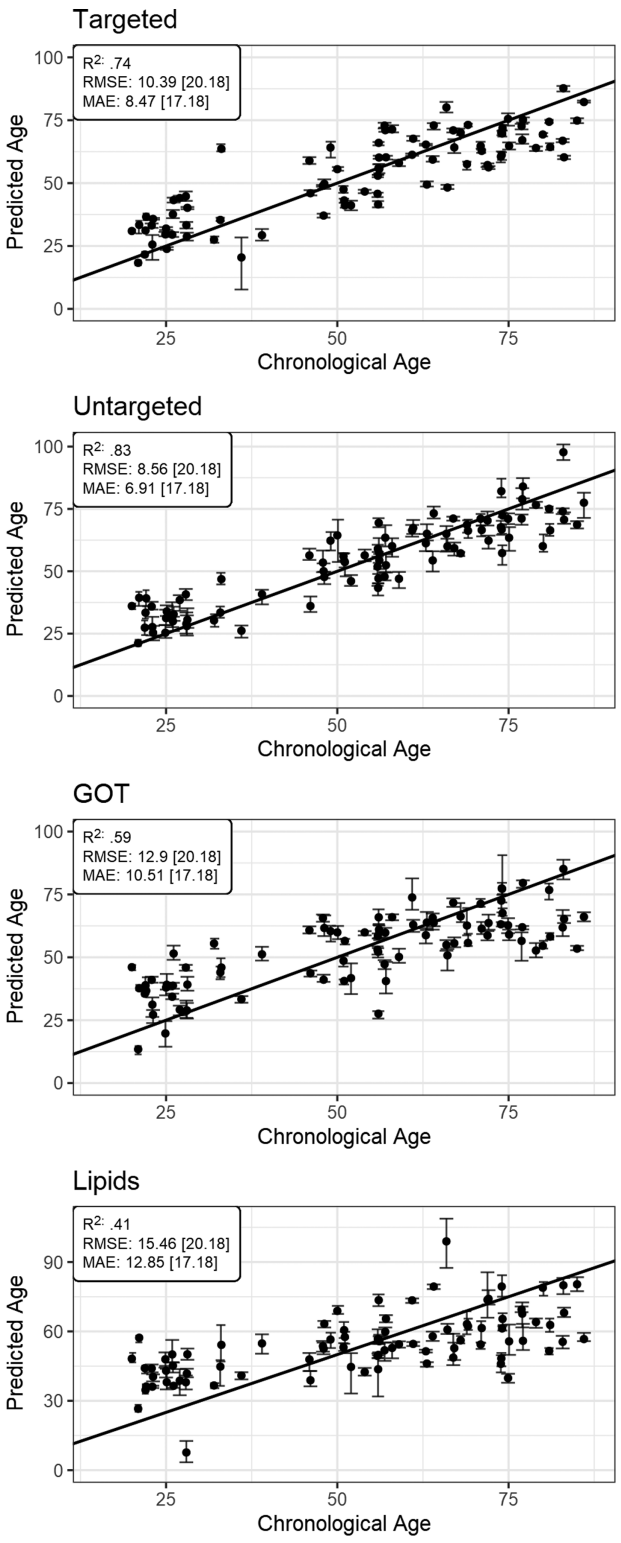

Figure 3. Predicted vs chronological age for controls in each profile, withR2, RMSE, and MAE reported. Numbers in parentheses represent the performance of the mean model for comparison. Points are the average of the predictions for the 5 imputations to estimate missing values, with the error bars representing the most extreme predicted values from the imputations. We also include the y = x line. Points above the line correspond to overestimates of a subject's chronological age, while points below the line correspond to underestimates. GOT = globally optimized targeted; MAE = mean absolute error; RMSE = root mean squared error.

with the most missingness in the lipidomic data. To quantify the relationship between age and missingness on aggregate, we fit an elastic net regression model where we replace metabolite concentrations with 0 if the value is missing, and 1 otherwise. For this model, we report a root mean squared error (RMSE) of 14.8, mean absolute error (MAE) of 12.4, and an R^2 of .47 in the untargeted data set (Supplementary Figure 4).

To characterize the relationship between age and missingness for individual lipids, we separate subjects into 3 equally spaced age groups and perform Pearson chi-squared tests on each lipid, testing the null hypothesis that the proportion of missingness is the same within each cohort. These tests found 6 lipids with FDR <.05, all of which are triglycerides (Supplementary Figure 4), which represents a 2.2-fold enrichment over what would be expected if 6 lipids were selected by chance. More broadly, our results demonstrate the importance of examining differential missingness across groups in the "omics" of aging.

Predictive Models for Age

We train separate age prediction models on each of our 4 data sets and found a MAE ranging from ~7 to 13 years, with the lowest error from the untargeted data set. Figure 2 depicts chronological age against the predicted age for each of these profiles, while Supplementary Figure 5 displays the same figure for the model combining all 4 profiles. The predictions using each of the 4 profiles (after detrending for chronological age) report Spearman rank-order correlations between .1 and .54, with the predictions from the lipid profiles reporting the lowest correlations (Supplementary Table 1). As a reference point, the intercept-only model, which predicts the mean age for every subject, yields an RMSE of 20.18 and MAE of 17.18. The average value of the penalty parameter (λ) across the leave-one-out prediction models is 1.40 for models fit using the untargeted profile, 1.49 for the targeted profile, 1.54 for the lipid profiles, and 3.06 for the GOT profile.

Additionally, models are fit on 3 variations of the untargeted profile to test the effects of drift correction in data preprocessing, to verify the modeling decision to keep sex exempt from penalization, and to test a biological age hypothesis. First, we verify that the drift correction procedure described in the "Data preprocessing" section improves out-of-sample predictions for age compared to using the uncorrected metabolite intensities in the untargeted profile, reporting an RMSE of 10.3, MAE of 8.4, and R^2 of .75. Compared to the performance of the models fit after drift correction (Figure 2), this represents an 18% increase in RMSE, 20% increase in MAE, and 10% reduction in R^2.

To verify the decision to exclude sex from regularization, models for male and female participants were fit separately on the untargeted profile. We obtain RMSEs of 11.1 and 12.8, MAEs of 9.5 and 9.8, and R^2 of .7 and .6, for male and female subjects, respectively. There are no metabolites consistently shared between the separate sex models among the 5 missing data imputed data sets, possibly indicating that the separate sex models are picking up different signals in the metabolome. Performing the same procedure on random samples of controls of the same sizes as the separate sex models (44 male subjects and 41 female subjects), we obtain similar predictive accuracy: RMSEs of 10.7 and 12.64, MAEs of 8.7 and 10.3, and R^2 of .68 and .54, for models fit using 44 and 41 subjects, respectively.

A previous epigenetic clock study based on 353 CpG sites tends to overpredict the age in a cohort of subjects with PD (43). To test whether our metabolomic age prediction model exhibits similar behavior, we obtained age predictions for a cohort of 57 AD and 56 PD subjects. Figure 4 displays the predictions of an age-matched control cohort alongside the AD/PD cohort for comparison. We find that the predictions for the AD/PD cohort are less accurate than those for the age-matched controls, with a 50% decrease in R^2, 30% increase in RMSE, and 18% increase in MAE. We also observe that within the Parkinson's cohort, the subject with the largest model error was one of the 6 pathogenic GBA carriers in our data set.

Age-Associated Metabolites

Elastic net automatically performs variable shrinkage by setting coefficients for nonpredictive metabolites to zero. Our models for the untargeted, targeted, and GOT-MS data sets typically included non-zero coefficients for between 30 and 40 metabolites. Supplementary Figure 6 shows summary statistics for the behavior of these models across the imputations and leave-one-out modeling procedure.

The models applied to the lipidomic data tended to include fewer features than the metabolite data sets, with an average of 15 retained across the 5 imputations. Table 1 lists the targeted metabolites and lipids that appeared in models fit on all 85 control subjects. As a summary, we find that in the model formed using the targeted profile, increased concentrations of xanthine, kynurenine, carnitine, hydroxy-indoleacetic acid (HIAA), and cystine were the largest contributors to higher age predictions, while increased concentrations

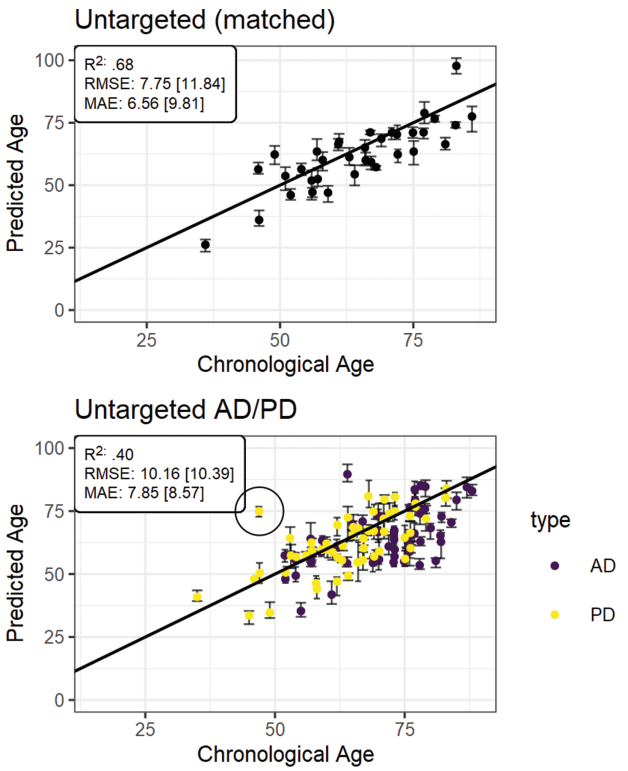

Figure 4. The leave-one-out performance of the untargeted model only looking at matched controls (top) and the performance of the model on the AD/PD cohort (bottom). The circled point is the greatest outlier (in terms of RMSE) for PD. Numbers in parentheses refer to the performance of the model which predicts the mean age for each subject. AD = Alzheimer's disease; MAE = mean absolute error; PD = Parkinson's disease; RMSE = root mean squared error. Full color version is available within the online issue.

of 4-aminobutyric acid, serine, and uridine were the largest contributors to lower age predictions. In the lipidome model, increased concentrations of sphingomyelin (18:1) [SM(18:1)], SM(16:0), triacylglycerol 52:2-FA 18:1, dihydroceramide (24:1), free fatty acids (24:0), and SM(14:0) were the largest contributors to higher age predictions, while increased concentrations of hexosylceramide (24:1) and phosphatidylethanolamine (O18:0/22:4) were the largest contributors to lower age predictions. Because the features were standardized prior to fitting the models, the listed coefficients represent the expected difference in age prediction resulting from a 1 *SD* increase in concentration. Note, however, that because our models are multivariate, these coefficients represent change in expected age given all of the other covariates, and should therefore not be interpreted unconditionally.

Pathway Analysis and Set Enrichment Analysis

While the untargeted data set yielded the smallest predictive error among all sets, the metabolite identities are unknown, making biological interpretation difficult. At FDR <.05, Mummichog identified the carnitine shuttle, starch and sucrose metabolism, putative anti-inflammatory metabolites formation from eicosapentaenoic acid, and biopterin metabolism pathways for the positive mode, and the glyoxylate and dicarboxylate metabolism pathway for the negative mode as associated with age (Supplementary Figure 7). The empirical compounds associated with the carnitine shuttle are displayed along with their marginal relationship with age in Figure 5.

We also use the targeted data to identify metabolomic pathways with an overrepresentation of metabolites that are significantly associated with age by MSEA (41). Using the Small Molecule Pathways Database (SMPDB) (44) or MetaboAnalyst's CSF disease-associated library, MSEA finds no associations below an FDR threshold of .05. However, our list of significant metabolites has large overlap with the tryptophan and tyrosine metabolism set in the SMPDB (Figure 5; Supplementary Table 2).

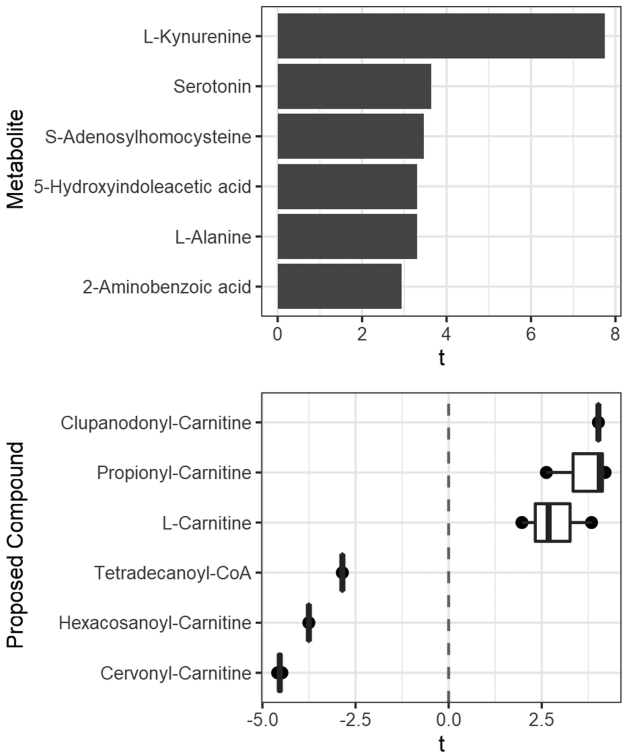

Figure 5. The marginal linear relationships between chronological age and the targeted metabolites overlapping the tryptophan metabolism pathway from the Small Molecule Pathways Database (top) and the untargeted metabolites identified by Mummichog to be associated with the carnitine shuttle (bottom). The *x*-axis contains the standardized relationship between metabolite concentration and age, expressed as a *t* statistic. Boxplots display the distribution of the statistic across untargeted metabolites in the case when Mummichog identifies the proposed compound in more than one metabolite.

Discussion

Two seminal papers authored by Horvath (8) and Hannum et al. (9) on epigenetic clocks inspired investigation into an assortment of biological clocks (5), including the metabolomic clock proposed by Rist et al. (16) and Robinson et al. (15). In this paper, we present predictive models for age using CSF metabolomic data. We find that metabolite and lipid data are generally able to predict chronological age within approximately 10 years. For the profiles where feature identity is known (targeted metabolomics and lipidomics), the model coefficients are reported to indicate which predictors are driving the model. We also find that younger subjects tended to have the most missing data in their metabolome profiles, which is potentially explained by the finding that CSF total protein increases with age (45). To our knowledge, this work represents the first metabolomic and lipidomic clock using mass spectrometry analysis of CSF.

Biological Interpretation of Targeted Multivariate Analysis

In the age prediction model built on the targeted metabolomic profile, we find that several of the metabolites driving the predictions (listed in Table 1) have been found to be associated with age. Consistent with our results, Blennow et al. (46) find a positive correlation between 5-HIAA and age in human CSF, and Johnson et al. (14) report negative correlation between serine and age in human plasma.

In the age prediction model fit on the lipidomic profile, SMs are the largest drivers of age predictions, as 3 SMs appear with positive associations with age. Collino et al. (47) find that SM(16:0) concentration increased with age (although other SM species exhibited the opposite behavior).

Biological Interpretation of Pathways Analysis

The pathways identified by Mummichog from the untargeted profile are consistent with the existing literature on age-related changes in the metabolome. The carnitine shuttle has previously been identified to be associated with age in humans (13). Because the carnitine shuttle is responsible for transportation of fatty acids to the mitochondria, these results might suggest an association between carnitine and age-related changes in the regulation of cellular energy (48). Komori et al. (49) find a relationship between the biopterin metabolism and age using CSF samples. Vitamin E metabolism has been reported to be associated with age by Robinson et al. (15) in a metabolomic aging analysis. In a study of mice, Ivanisevic et al. (21) identify purine to be related with aging in the brain. Several of these pathways, the carnitine shuttle, Vitamin E metabolism, and tryptophan metabolism observed here, have also been associated with frailty (50).

MSEA does not find any significant associations for targeted univariate analysis at FDR <.05. However, the *p* values for this procedure are computed using overrepresentation analysis based on

the cumulative hypergeometric distribution to compare the input list of significant metabolites to the SMPDB. As such, our input list can contain all the metabolites in small SMPDB pathways (in our case, tyrosine and glutamate metabolism) and the FDR would still be above .05. For our exploratory purposes, these associations can still be valuable. In particular, an association between tyrosine and age has been found in female human subjects in a longitudinal metabolomic study (51), and tryptophan, which plays a role in the regulation of both neuronal activity and immune response, has been featured prominently in the literature on the aging metabolome (7,47,52). In our analysis, the 2 metabolites with the strongest association with age within the tryptophan pathway are kynurenine and serotonin, which might suggest that both immune regulation and neuropsychology are important contributors to age-related changes in the metabolome (53).

Interpreting Aging Models Across Profiles

We observed considerable differences in results between profiles. These are likely caused by a combination of true differences between the data sets, regularization bias, and varying amounts of preprocessing needed to run the models. For instance, the lipid model was most sensitive to the effects of missing data, as 80% of the data set was missing. Many features were excluded (those with >50% missingness), and the rest were affected by missing data imputation. This lends itself to regularization bias, as the amount of signal in a feature is dampened due to the imputed values. Supplementary Discussion 1 showcases 2 attempts at reducing regularization bias in the modeling procedure: preselecting features using a univariate correlation threshold and postselection by elastic net (Supplementary Figures 8-10). Neither method significantly improved results, but we find that postselection inference can be useful when the researcher expects a small number of features to explain a large percentage of the variation in age (although it is likely invalid in the small data setting), while feature reduction by univariate correlation can be useful when the researcher expects the predictive power of the metabolites to be largely independent of one another. The comparable errors between these models suggests the presence of many equally performant age prediction models of the metabolome and can possibly explain the similar performance of the models fit on each sex separately despite having no consistently shared metabolites.

Consideration should also be made for computational efficiency. An empirical prior is used to speed up the missing data imputation process but comes at the cost of shrinking covariances between predictors. This tradeoff is necessary because of the complexity of the leave-one-out prediction procedure, which is used to approximate out-of-sample performance in a small data setting. A separate validation set would have eliminated the need for these simplifications altogether.

We also note that our data represent a small cross-sectional sample, which comes with inherent limitations when studying a temporal variable. While the control subjects used in this study were medically healthy and cognitively normal volunteers, it is possible that unmeasured variables influence the results. One potential confounder of concern is diet, as it can both systematically differ by age and impact metabolite concentrations. As such, the pathways identified by the univariate analysis of the untargeted and targeted profiles are not attempts to claim understanding of the biological mechanisms behind aging, and the model predictions should not be taken as a measure of biological age. Rather, this analysis is most helpful when considered in the context of existing work, and as a starting point for more focused future work.

Alzheimer's, Parkinson's, and the Biological Clock

In the application of our age prediction models to the AD/PD subjects, the performance is not much better than the intercept-only model, and there is no evidence of the consistent overprediction present in Horvath and Ritz (43). On the contrary, our models exhibit a pattern of underprediction on these subjects. This behavior could be due to differences in the metabolome of the AD/PD cohort which cannot be attributed to age. However, this behavior could also be unrelated to disease status, as a pattern of underestimating the age of older subjects appears in the age-matched control cohort in Figure 4, and is also a pattern observed in epigenetic clocks (54). This underprediction could be due to regularization bias induced by our modeling procedure, but it is also possible that metabolite concentrations become less informative of age as time progresses. Supplementary Figure 11 flips our regression around, regressing individual metabolite concentrations by age, and demonstrating a possible leveling off of metabolite concentration at older ages.

A limitation to this analysis is that the models used to generate predictions for the AD/PD cohort could not be properly validated on control subjects. This is because we opted to form new elastic net models using all 85 of the control subjects, rather than applying each of the leave-one-out prediction models. While the model coefficients are similar to the leave-one-out models used to form the predictions shown in Figure 2, a separate validation cohort of control subjects is needed for a fair comparison to the strictly out-of-sample predictions on the AD/PD subjects.

Conclusion

A metabolomic analysis of the CSF has the potential to capture physiological variation that changes slowly over time, and that reflects variation in the central nervous system. However, it is important to keep in mind that the methods and results are limited due to the small size and cross-sectional nature of our data, which makes the "metabolomic age" established in this paper less reliable for use as an aging biomarker, as metrics akin to epigenetic "age acceleration" are impacted by these constraints (43). However, these limitations can be common when working with data from such invasive procedures as the extraction of CSF samples. We hope that this paper serves to illustrate methods that can be used to extract maximal information in these data sets with small sample size and large features, as well as showcase the predictive power and usefulness of such studies. We also hope that this work will motivate larger studies and analysis of longitudinal cohorts, with the goal of developing a more robust aging model using the metabolome.

Supplementary Material

Supplementary data are available at *The Journals of Gerontology, Series A: Biological Sciences and Medical Sciences* online.

Funding

This work was supported by the National Institutes of Health (grant numbers P50 NS062684, P50 AG05136, and S10 OD021562), the Department of Veterans Affairs (grant number 101 CX001702), the Veterans Affairs Northwest Mental Illness Research, Education, and Clinical Center, and an anonymous foundation. D.E.L.P. was supported by the National Institutes of

Health (grant numbers R01 AG049494 and R01 AG057330). D.E.L.P. and A.F. were supported by the National Institutes of Health (grant number R03 CA211160). Metabolomic assays were supported in part by the UW Nathan Shock Center of Excellent for the Biology of Aging grant P30 AG013280.

Conflict of Interest

None declared.

References

1. Christensen K, Vaupel JW. Determinants of longevity: genetic, environmental and medical factors. *J Intern Med*. 1996;240(6):333–341. doi:10.1046/j.1365-2796.1996.d01-2853.x
2. Deelen J, Beekman M, Uh HW, et al. Genome-wide association study identifies a single major locus contributing to survival into old age; the APOE locus revisited. *Aging Cell*. 2011;10(4):686–698. doi:10.1111/j.1474-9726.2011.00705.x
3. Walter S, Atzmon G, Demerath EW, et al. A genome-wide association study of aging. *Neurobiol Aging*. 2011;32(11):2109.e15–2109.e28. doi:10.1016/j.neurobiolaging.2011.05.026
4. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature*. 2009;461(7265):747–753. doi:10.1038/nature08494
5. Kudryashova KS, Burka K, Kulaga AY, Vorobyeva NS, Kennedy BK. Aging biomarkers: from functional tests to multi-omics approaches. *Proteomics*. 2020;20(5–6):e1900408. doi:10.1002/pmic.201900408
6. Jin K, Wilson KA, Beck JN, et al. Genetic and metabolomic architecture of variation in diet restriction-mediated lifespan extension in *Drosophila*. *PLoS Genet*. 2020;16(7):e1008835. doi:10.1371/journal.pgen.1008835
7. Laye MJ, Tran V, Jones DP, Kapahi P, Promislow DE. The effects of age and dietary restriction on the tissue-specific metabolome of *Drosophila*. *Aging Cell*. 2015;14(5):797–808. doi:10.1111/acel.12358
8. Horvath S. DNA methylation age of human tissues and cell types. *Genome Biol*. 2013;14(10):R115. doi:10.1186/gb-2013-14-10-r115
9. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Mol Cell*. 2013;49(2):359–367. doi:10.1016/j.molcel.2012.10.016
10. Chen BH, Marioni RE, Colicino E, et al. DNA methylation-based measures of biological age: meta-analysis predicting time to death. *Aging (Albany NY)*. 2016;8(9):1844–1865. doi:10.18632/aging.101020
11. Harrison BR, Wang L, Gajda E, et al. The metabolome as a link in the genotype–phenotype map for peroxide resistance in the fruit fly, *Drosophila melanogaster*. *BMC Genomics*. 2020;21(1):341. doi:10.1186/s12864-020-6739-1
12. Deelen J, Kettunen J, Fischer K, et al. A metabolic profile of all-cause mortality risk identified in an observational study of 44,168 individuals. *Nat Commun*. 2019;10(1):3346. doi:10.1038/s41467-019-11311-9
13. Lawton KA, Berger A, Mitchell M, et al. Analysis of the adult human plasma metabolome. *Pharmacogenomics*. 2008;9(4):383–397. doi:10.2217/14622416.9.4.383
14. Johnson LC, Parker K, Aguirre BF, et al. The plasma metabolome as a predictor of biological aging in humans. *Geroscience*. 2019;41(6):895–906. doi:10.1007/s11357-019-00123-w
15. Robinson O, Chadeau Hyam M, Karaman I, et al. Determinants of accelerated metabolomic and epigenetic aging in a UK cohort. *Aging Cell*. 2020;19(6):e13149. doi:10.1111/acel.13149
16. Rist MJ, Roth A, Frommherz L, et al. Metabolite patterns predicting sex and age in participants of the Karlsruhe Metabolomics and Nutrition (KarMeN) study. *PLoS One*. 2017;12(8):e0183228. doi:10.1371/journal.pone.0183228
17. van den Akker EB, Trompet S, Barkey Wolf JJH, et al. Metabolic age based on the BBMRI-NL ^{1}H-NMR metabolomics repository as biomarker of age-related disease. *Circ Genom Precis Med*. 2020;13(5):541–547. doi:10.1161/CIRCGEN.119.002610
18. Gu H, Pan Z, Xi B, et al. ^{1}H NMR metabolomics study of age profiling in children. *NMR Biomed*. 2009;22(8):826–833. doi:10.1002/nbm.1395
19. Wishart DS, Lewis MJ, Morrissey JA, et al. The human cerebrospinal fluid metabolome. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2008;871(2):164–173. doi:10.1016/j.jchromb.2008.05.001
20. Ma S, Yim SH, Lee SG, et al. Organization of the mammalian metabolome according to organ function, lineage specialization, and longevity. *Cell Metab*. 2015;22(2):332–343. doi:10.1016/j.cmet.2015.07.005
21. Ivanisevic J, Stauch KL, Petrascheck M, et al. Metabolic drift in the aging brain. *Aging (Albany NY)*. 2016;8(5):1000–1020. doi:10.18632/aging.100961
22. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Roy Stat Soc B*. 2005;67(2):301–320. doi:10.1111/j.1467-9868.2005.00503.x
23. Thompson MJ, vonHoldt B, Horvath S, Pellegrini M. An epigenetic aging clock for dogs and wolves. *Aging (Albany NY)*. 2017;9(3):1055–1068. doi:10.18632/aging.101211
24. Mata IF, Leverenz JB, Weintraub D, et al. GBA variants are associated with a distinct pattern of cognitive deficits in Parkinson's disease. *Mov Disord*. 2016;31(1):95–102. doi:10.1002/mds.26359
25. Mata IF, Leverenz JB, Weintraub D, et al. APOE, MAPT, and SNCA genes and cognitive performance in Parkinson disease. *JAMA Neurol*. 2014;71(11):1405–1412. doi:10.1001/jamaneurol.2014.1455
26. Kim HM, Nazor C, Zabetian CP, et al. Prediction of cognitive progression in Parkinson's disease using three cognitive screening measures. *Clin Park Relat Disord*. 2019;1:91–97. doi:10.1016/j.prdoa.2019.08.006
27. Cholerton BA, Zabetian CP, Quinn JF, et al. Pacific Northwest Udall Center of Excellence Clinical Consortium: study design and baseline cohort characteristics. *J Parkinsons Dis*. 2013;3(2):205–214. doi:10.3233/JPD-130189
28. Gibb WR, Lees AJ. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *J Neurol Neurosurg Psychiatry*. 1988;51(6):745–752. doi:10.1136/jnnp.51.6.745
29. Zhang X, Dong J, Raftery D. Five easy metrics of data quality for LC-MS-based global metabolomics. *Anal Chem*. 2020;92(19):12925–12933. doi:10.1021/acs.analchem.0c01493
30. Gu H, Zhang P, Zhu J, Raftery D. Globally optimized targeted mass spectrometry: reliable metabolomics analysis with broad coverage. *Anal Chem*. 2015;87(24):12355–12362. doi:10.1021/acs.analchem.5b03812
31. Shi X, Wang S, Jasbi P, et al. Database-assisted globally optimized targeted mass spectrometry (dGOT-MS): broad and reliable metabolomics analysis with enhanced identification. *Anal Chem*. 2019;91(21):13737–13745. doi:10.1021/acs.analchem.9b03107
32. Zhong F, Xu M, Zhu J. Development and application of time staggered/mass staggered-globally optimized targeted mass spectrometry. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2019;1120:80–88. doi:10.1016/j.jchromb.2019.04.051
33. Hanson AJ, Banks WA, Bettcher LF, Pepin R, Raftery D, Craft S. Cerebrospinal fluid lipidomics: effects of an intravenous triglyceride infusion and apoE status. *Metabolomics*. 2019;16(1):6. doi:10.1007/s11306-019-1627-x
34. Morris C. A finite selection model for experimental design of the health insurance study. *J Econ*. 1979;11(1):43–61. doi:10.1016/0304-4076(79)90053-8
35. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016;785–794. doi:10.1145/2939672.2939785
36. Honaker J, King G, Blackwell M. Amelia II: a program for missing data. *J Stat Soft*. 2011;45(7):1-47. doi:10.18637/jss.v045.i07
37. Thévenot EA, Roux A, Xu Y, Ezan E, Junot C. Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses. *J Proteome Res*. 2015;14(8):3322–3335. doi:10.1021/acs.jproteome.5b00354
38. Jové M, Maté I, Naudí A, et al. Human aging is a metabolome-related matter of gender. *J Gerontol A Biol Sci Med Sci*. 2016;71(5):578–585. doi:10.1093/gerona/glv074

39. Saar-Tsechansky M, Provost F. Handling missing values when applying classification models. *J Mach Learn Res*. 2007;8:1623–1657. doi:10.5555/1314498.1314553
40. Li S, Park Y, Duraisingham S, et al. Predicting network activity from high throughput metabolomics. *PLoS Comput Biol*. 2013;9(7):e1003123. doi:10.1371/journal.pcbi.1003123
41. Xia J, Wishart DS. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res*. 2010;38(Web Server issue):W71–W77. doi:10.1093/nar/gkq329
42. Pang Z, Chong J, Li S, Xia J. MetaboAnalystR 3.0: toward an optimized workflow for global metabolomics. *Metabolites*. 2020;10(5):186. doi:10.3390/metabo10050186
43. Horvath S, Ritz BR. Increased epigenetic age and granulocyte counts in the blood of Parkinson's disease patients. *Aging (Albany NY)*. 2015;7(12):1130–1142. doi:10.18632/aging.100859
44. Frolkis A, Knox C, Lim E, et al. SMPDB: the Small Molecule Pathway Database. *Nucleic Acids Res*. 2010;38(Database issue):D480–D487. doi:10.1093/nar/gkp1002
45. Breiner A, Moher D, Brooks J, et al. Adult CSF total protein upper reference limits should be age-partitioned and significantly higher than 0.45 g/L: a systematic review. *J Neurol*. 2019;266(3):616–624. doi:10.1007/s00415-018-09174-z
46. Blennow K, Wallin A, Gottfries C, et al. Cerebrospinal fluid monoamine metabolites in 114 healthy individuals 18–88 years of age. *Eur Neuropsychopharmacol*. 1993;3(1):55–61. doi:10.1016/0924-977X(93)90295-W
47. Collino S, Montoliu I, Martin FP, et al. Metabolic signatures of extreme longevity in northern Italian centenarians reveal a complex remodeling of lipids, amino acids, and gut microbiota metabolism. *PLoS One*. 2013;8(3):e56564. doi:10.1371/journal.pone.0056564
48. Flanagan JL, Simmons PA, Vehige J, Willcox MD, Garrett Q. Role of carnitine in disease. *Nutr Metab (Lond)*. 2010;7:30. doi:10.1186/1743-7075-7-30
49. Komori H, Matsuishi T, Yamada S, Ueda N, Yamashita Y, Kato H. Effect of age on cerebrospinal fluid levels of metabolites of biopterin and biogenic amines. *Acta Paediatrica*. 1999;88(12):1344–1347. doi:10.1111/j.1651–2227.1999.tb01048.x
50. Rattray NJW, Trivedi DK, Xu Y, et al. Metabolic dysregulation in vitamin E and carnitine shuttle energy mechanisms associate with human frailty. *Nat Commun*. 2019;10(1):5027. doi:10.1038/s41467-019-12716-2
51. Chak CM, Lacruz ME, Adam J, et al. Ageing investigation using two-time-point metabolomics data from KORA and CARLA studies. *Metabolites*. 2019;9(3):44. doi:10.3390/metabo9030044
52. van der Goot AT, Nollen EA. Tryptophan metabolism: entering the field of aging and age-related pathologies. *Trends Mol Med*. 2013;19(6):336–344. doi:10.1016/j.molmed.2013.02.007
53. Hestad KA, Engedal K, Whist JE, Farup PG. The relationships among tryptophan, kynurenine, indoleamine 2,3-dioxygenase, depression, and neuropsychological performance. *Front Psychol*. 2017;8:1561. doi:10.3389/fpsyg.2017.01561
54. El Khoury LY, Gorrie-Stone T, Smart M, et al. Systematic underestimation of the epigenetic clock and age acceleration in older subjects. *Genome Biol*. 2019;20(1):283. doi:10.1186/s13059-019-1810-4