AgeML: Age modeling with Machine Learning

Jorge Garcia Condado, Iñigo Tellaetxe Elorriaga, Jesus M. Cortes, Asier Erramuzpe

***Abstract*— An approach to age modeling involves the supervised prediction of age using machine learning from subject features. The derived age metrics are used to study the relationship between healthy and pathological aging in multiple body systems, as well as the interactions between them. We lack a standard for this type of age modeling. In this work we developed AgeML, an OpenSource software for age-prediction from any type of tabular clinical data following well-established and tested methodologies. The objective is to set standards for reproducibility and standardization of reporting in supervised age modeling tasks. AgeML does age modeling, calculates age deltas, the difference between predicted and chronological age, measures correlations between age deltas and factors, visualizes differences in age deltas of different clinical populations and classifies clinical populations based on age deltas. With this software we are able to reproduce published work and unveil novel relationships between body organs and polygenetic risk scores. AgeML is age modeling made easy for standardization and reproducibility.**

***Index Terms*— Brain Age, Aging, Machine Learning, Open Source, Standardization, Reproducibility, Polygenetic Risk Scores**

I. Introduction

Age modeling can be implemented using supervised machine learning regression in which clinical, behavioral, cognitive or brain/body morphological features of a subject are used to predict their chronological age [1]. Age models can be built using body organ specific features to create organ specific models, or organ ages [1]. Models have been built for many organs such as the brain [2]–[4], heart [5], eye [6], but also from cognitive performance [7]. These models can integrate information from a single source of data or from multiple sources, thereby facilitating the development of multi-modal age models. The emergence of such age models is attributed to the accessibility of large-scale healthcare databases in the recent decade, exemplified by the UKBiobank [8].

Researchers are focused on elucidating the age gap or delta, the discrepancy between the predicted age and the chronological age of a subject. Age models are trained on two different types of populations: healthy participants devoid of chronic disorders, or on a sample representative of the general population. These deltas, representing the residuals in the regression task, have been shown to correlate with lifestyle factors. Participants engaging in unhealthy lifestyle practices, including smoking, excessive alcohol consumption, or sedentary behavior, exhibit elevated deltas [1]. Age deltas have also been recently associated to specific genetic mutations [9]. The variation in age deltas across organs at different time points, and their multi-variate integration, can help us understand large-scale physiological multi-organ networks [1], of special relevance for the field of network physiology [10]. Participants diagnosed with diseases exhibit higher deltas in specific organs; for example, variations in brain age deltas can distinguish individuals with a neurodegenerative disease such as Alzheimer's disease from healthy controls [11].

These deltas have two key applications. First, they can be used for screening and identifying outliers that may require further clinical follow-up. For instance, previous studies have demonstrated that brain age deltas are correlated with the time to conversion for Alzheimer's disease, providing a quantifiable measure of risk [7]. Second, deltas can help quantify deviations from healthy aging and be included as covariates in research, allowing for the separation of aging-related effects from pathological processes [12], [13]. However, from the practical perspective of implementing models and calculating deltas, there still exist many problems and questions in this emerging multidisciplinary field, particularly concerning the implementation, standardization, and reproducibility across institutions.

AgeML is an OpenSource Python package that makes age modeling with machine learning easy. AgeML has three main goals: standardization, reproducibility and easy access. Standardization is the first element necessary to enable the comparison of studies. Reproducibility is crucial for fostering trust within the community and ensuring the consistency of results. It is vital to provide an easy start for individuals interested in age modeling, particularly for clinicians with potentially less theoretical backgrounds in machine learning. We believe that AgeML could serve as a foundational platform

The project that gave rise to these results received the support of a fellowship from "la Caixa" Foundation (ID 100010434) to JGC. The fellowship code is LCF/BQ/DI21/11860030. JMC is funded by Ikerbasque: The Basque Foundation for Science, the Department of Economic Development and Infrastructure of the Basque Country (Elkartek Program, grant KK-2021-00009), and the Health Department of the Basque Country (grants 2022111031 and 2023111002). AE is supported by Ikerbasque, the Basque Foundation for Science, and the Spanish Ministry of Science and Innovation, grant RYC2021-032390-I. *(JCG and ITE contributed equally to this work) (Corresponding author: AE)*

JGC and ITE are with the Computational Neuroimaging Lab, Biobizkaia Health Research Institute, Barakaldo, Spain, 48903 and the Biomedical Research Doctorate Program, University of the Basque Country, Leioa, Spain, 48940.

JMC is with the Computational Neuroimaging Lab, Biobizkaia Health Research Institute, Barakaldo, Spain, 48903 and the Department of Cell Biology and Histology, University of the Basque Country, Leioa, Spain, 48940 and IKERBASQUE, the Basque Foundation for Science, Bilbao, Spain, 48009.

AE is with the Computational Neuroimaging Lab, Biobizkaia Health Research Institute, Barakaldo, Spain, 48903 and IKERBASQUE, the Basque Foundation for Science, Bilbao, Spain, 48009. (email: asier.erramuzpe@gmail.com)

for addressing the research questions within the field, led by developers and researchers themselves.

II. Related Work

Previous works on machine learning-based age modeling have focused on the derived biomarker called the age delta, the difference between the chronological and predicted age. Specifically, understanding if these deltas can be used as a proxy for deviations from a healthy aging trajectory. Research has focused on whether these deviations come from a sudden change, an increase in the delta at a specific time point, or an accumulation of small changes leading to a higher delta over time [14]. This age delta can be used for a classification task downstream. A debate exists regarding whether research should prioritize minimizing the error in age estimation tasks or, alternatively, focus on developing models that lead to deltas that are generalizable for subsequent classification tasks [15]–[17], ie., differentiating between individuals at higher risk of developing a pathological condition. There is also large variability between reported average deltas of disease groups across different studies [18]. Caution is essential in selecting training data, as it often comes from white populations. This could lead to possible biases, such as an increase in age prediction error in controls. This has been observed when applying models trained on white populations to non-white populations [19], [20]. Similar biases have also been found when applying a model across different age ranges than those it was trained in [21]. To tackle all these open questions, standardization in age modeling procedures is needed.

A previous effort aimed at developing a tool specific for brain age modeling is the Brain Age Standardized Evaluation (BASE) initiative [22]. BASE focuses on age modeling from brain magnetic resonance images (MRI), encompassing various scenarios such as scanning at multi-site, generalizing to unseen sites, test-retest, and longitudinal data. It is well known that MRI images can vary greatly between sites and even between images of the same subject taken with the same machine [23]–[25]. It also employs an evaluation protocol involving repeated training of brain age prediction models, assessing performance metrics for accuracy and robustness, to ultimately ensure reproducibility.

BASE is limited to working with neuroimaging data and produces BrainAge models, whereas AgeML operates with tabular data, allowing it to generate models for various organs beyond the brain. For example, while AgeML can produce BrainAge models from brain volumetric tabular data, it can also create models for the heart using routine clinical heart evaluations. Importantly, images are often more expensive, time-consuming, and less accessible compared to tabular data from routine clinical practice due to anonymity concerns. Additionally, images typically require extensive preprocessing compared to tabular data. Age models based on imaging also demand more complex architectures and larger datasets, making them harder to train compared to tabular data-based models. However, one limitation of AgeML is the need to convert all data into tabular form, which could potentially result in a loss of information.

AgeML expands upon many of the functionalities of BASE, enabling multi-modal inputs of different organ features, which aligns with a more contemporary approach of system medicine viewing the body as a network of interacting organs [10]. This makes AgeML more accessible to a broader research community. Moreover, AgeML is open-source developed with transparency to ensure that standards are led by the community. Additionally, AgeML is specifically focused on analysis procedures for understanding clinical populations, thereby offering added value to the medical research community. There is a need to understand model variability [17] and for this to be done properly we need to ensure standardization and reproducibility across all different organ age models for different pathologies.

III. Data

The UKBiobank database [8] was used for this project (project ID 79105). Overall, the UKBiobank consists of a total of 502,504 individuals, of which 229,122 are males. These participants, aged between 37 and 73 years during recruitment (2006–2010), underwent thorough assessments, including questionnaires, physical examinations, blood and urine sample analyses, and genome-wide genotyping. The extensive data collection occurred across 22 assessment centers located throughout the United Kingdom.

Body phenotypes: We created age models using 78 different body phenotypes, which align with those in the work of Ye Tian et al. 2023 [1]. A breakdown of these body phenotypes can be found in Supplementary Table S1 of Ye Tian et al. 2023 [1]. The different body phenotypes were grouped into seven distinct organ systems: Musculoskeletal, Cardiovascular, Pulmonary, Immune, Renal, Hepatic, and Metabolic. The organs each phenotype belongs to can be found under the Body age phenotypes subsection of Methods in Ye Tian et al. 2023 [1]. These groups were used to create distinct age models for each organ system. The 78 distinct phenotypes were also considered as features to create a unified age model, referred to as Body.

Clinical categories: We established them using ICD-10, ICD-9 codes and self-reported pathologies, translated into clinical categories by using the Clinical Classifications Software Refined [26]. Data was used based only on first occurrence labels. The distinction between chronic and non-chronic participants was established using the Chronic Condition Indicator Refined [27], resulting in a healthy population of N=6,388. Eight clinical categories were finally selected, concerning the organ systems previously defined: blood diseases (BLD), circulatory system diseases (CIR), endocrine, nutritional and metabolic diseases (END), mental, behavioral and neurodevelopmental disorders (MBD), Neoplasms (NEO), diseases of the musculoskeletal system (MUS), diseases of the nervous system (NVS), and respiratory system diseases (RSP). Participants with multiple comorbidities were excluded to mitigate potential mixed effects.

Polygenetic Risk Scores (PRS): To investigate the relationship between organ deltas and genetics, if any, a factor correlation analysis was carried out with PRS derived from

the genetic data of the UKBiobank. Two types of PRS were used: Standard PRS and Enhanced PRS [28]. Standard PRS are sourced from external UKBiobank GWAS data, while Enhanced PRS are derived from a combination of external and internal UK Biobank data. Notably, Enhanced PRS encompass a larger number of categories (53) compared to Standard PRS (39). Standard PRS were available for 6,323 out of the 6,388 participants, whereas Enhanced PRS were only accessible for 1,234 of these individuals. Prior to conducting the factor correlation analysis, the PRS were adjusted for the covariates recruitment site, sex, and ethnicity.

Additional information regarding the total number of participants per category and their age distributions can be found in Table I.

TABLE I
AGE DISTRIBUTION SUMMARY OF THE CLINICAL CATEGORIES

CLINICAL CATEGORY	$\mu_{age} \pm \sigma_{age}$ (YEARS)	AGE RANGE (YEARS)	N (PARTICIPANTS)
CN	51.39 ± 7.31	[40-70]	6,388
CIR	56.00 ± 7.90	[40-70]	3,907
END	55.43 ± 7.78	[40-70]	1,616
MBD	50.86 ± 7.04	[40-70]	1,184
MUS	54.11 ± 7.67	[39-70]	5,603
NVS	53.10 ± 7.63	[39-70]	983
RSP	51.41 ± 7.57	[40-70]	4,376
NEO	56.09 ± 7.75	[40-70]	2,215
BLD	51.76 ± 7.33	[40-69]	338

CN: Healthty Controls, CIR: Circulatory system disease, MBD: neurodevelopmental disorders, MUS: Musculoskeletal system disorders, RSP: respiratory system diseases, NEO: neoplasms, BLD: blood diseases

IV. METHODS

AgeML is developed as a Python package readily available to install through PyPI [29]. An overview of all AgeML functionalities can be seen in Fig. 1. There are three main strategies to work with AgeML. The first one is a simple command line interface that can be launched after installation by using the command `ageml`. This interface interactively asks for the required individual inputs at each different stage. The second strategy is by using the four available commands, each for a different pipeline: `model_age` for age modeling, `factor_correlation` to study the relationships between deltas and clinical outcomes, `clinical_groups` to study the deltas of different clinical groups and `clinical_classify` to use deltas as predictors in a classification task. Lastly, `ageml` can be used as an importable Python library, allowing for the implementation of custom functionality. The Open Source code can be found in GitHub: https://github.com/compneurobilbao/ageml.

AgeML provides a comprehensive machine learning pipeline for age modeling with four core functions. The first module `model_age` (green panel in Fig. 1) predicts age using a features file with optional system-specific modeling. It allows for the training of models on healthy controls and outputting predicted, corrected ages, and the delta (difference between corrected and chronological age) for each of the uploaded groups. The second module `factor_correlation` (red panel in Fig. 1) analyzes Pearson correlations between selected numerical factors (e.g., exercise hours) and the age delta, with the option to apply covariate correction. This can be done separately for each clinical group; based only on the control group and then applied to the whole population; or based on the whole population and then applied to it. The third module `clinical_groups` (blue panel in Fig. 1) visualizes age deltas across different clinical groups using box plots and performs statistical tests, including FDR and Bonferroni-corrected T-tests, to assess group differences in age deltas. The fourth module `clinical_classify` (purple panel in Fig. 1) utilizes Logistic Regression with cross-validation to classify individuals into clinical categories based on age deltas. It supports system-specific and comprehensive models across multiple systems. This structured pipeline allows for flexible age prediction, factor correlation analysis, and clinical group differentiation based on machine learning methods. A toy dataset is provided and concrete examples of use and expected outputs can be found in the Tutorial.

A. Age modeling

Age modeling can be effectively applied through supervised machine learning regression, where subject features are employed to predict the known chronological age of the subject (Y), corresponding to the subject's birth date. Let Y denote the predicted age, and $\mathbf{X}$ represent the model features. The aim is to find $f()$, our age modeling pipeline, such that:

$$Y = f(\mathbf{X}). \tag{1}$$

The pipeline consists of three steps: a scaler, a feature extension step, and a model. The available scalers are all imported from the Scikit-learn Python library [30] and include the max absolute scaler, min-max scaler, normalizer, power transformer, quantile transformer, robust scaler, and standard scaler. For feature extension, the polynomial feature extension from Scikit-learn [30] is used. The various models implemented consist of Linear Regression, Ridge, LASSO, Support Vector Regression (SVR), Random Forests (RF) and XGBoost [31].

AgeML contains a features vs age component that offers a visual and statistical check to assess potential correlations between features and chronological age. By computing Pearson's correlation coefficients, it provides a quick sanity check on whether the features may be informative for age prediction. While Pearson's correlation only captures linear relationships, visual inspection of the data may reveal non-linear patterns, prompting users to explore models better suited for capturing such relationships. This step ensures a more comprehensive understanding of the feature set and its relevance for accurate age modeling.

The feature extension module is not used as a default. The feature extension is left for modeling non-linear interactions of variables that could help in the downstream age modeling task. Many body variables follow a u-shaped pattern with age increasing until mid age and then decreasing in old age or vice versa. For example, when doing brain age modeling

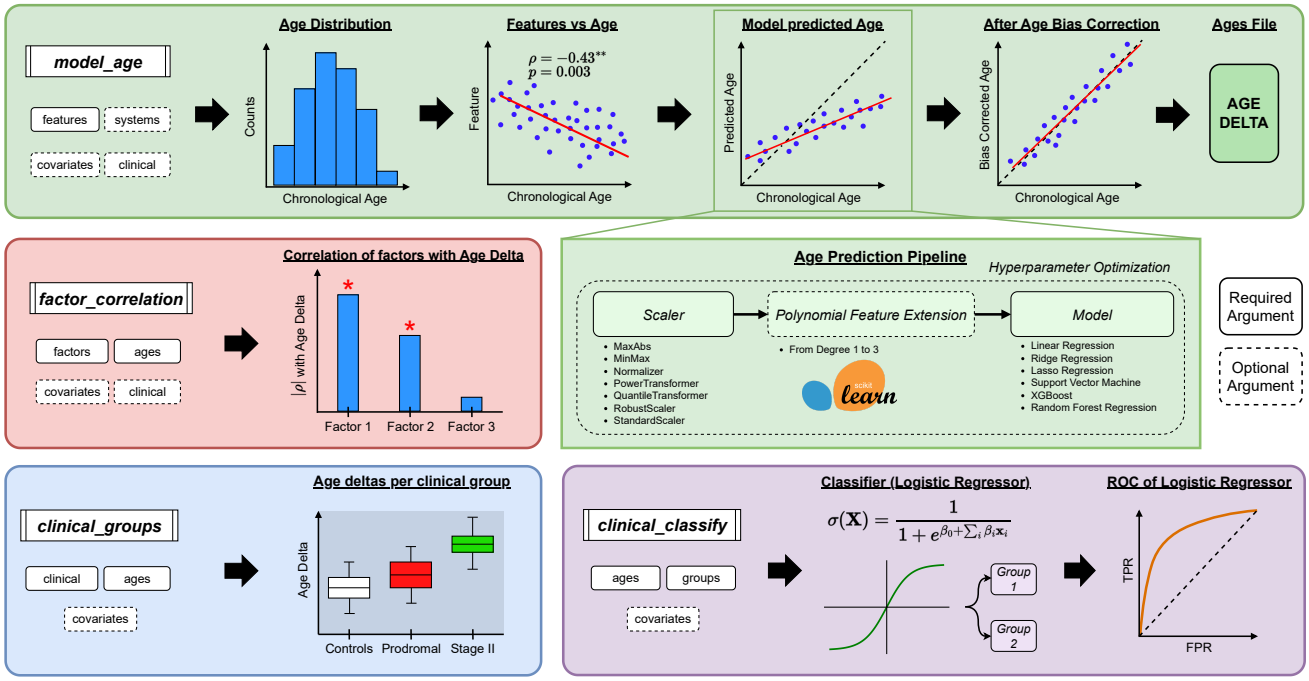

Fig. 1. **Age modeling with Machine Learning (AgeML), general overview, and major capabilities.** *Green panels:* the `model_age` module requires only a features file to operate, which should contain the chronological age of the subject and the set of features $\mathbf{x_i}$ for each individual. A system file consisting of sets of features (e.g.: a set that encompasses the musculoskeletal system and/or another set referring to renal features) can be provided to train an age prediction model for each specific system. For training a model that distinguishes participants based on a categorical value, a covariates file can be provided, enabling the performance of covariate correction. If a clinical file is provided indicating which participants are healthy controls, the age prediction model is exclusively trained using their features. This pipeline displays the age distribution of the healthy controls, the Pearson correlation ($\boldsymbol{\rho}$) and significance of each feature and chronological age (separated by system and covariate as required), and the predicted age as well as the bias corrected age. Finally, an ages file in CSV format is generated, incorporating the participant's chronological ages, their predicted uncorrected, corrected ages and the delta, the difference between corrected and chronological age. The age prediction pipeline includes the following components: a scaler selection step, an optional feature extension step, a model selection step, and the availability of optional hyperparameter optimization. *Red panel:* the `factor_correlations` module analyzes the Pearson correlation $\boldsymbol{\rho}$ strength and significance between the specified numerical factors (e.g.: weekly exercise hours, years of education) and the calculated age delta. Additionally, covariate correction can be applied, and this analysis can be performed for each clinical group as needed. *Blue panel:*, the `clinical_groups` module creates a box plot illustrating the age deltas for each clinical group, reporting the FDR and Bonferroni corrected T-test p-values and T-statistics of all groups. Although not depicted in the figure, the age distributions of all clinical groups are plotted, and a T-test is performed to assess group differences, reporting their p-values. *Purple panel:* The `clinical_classify` module employs Logistic Regression with *n*-fold cross-validation and a defined sigmoid threshold to differentiate between two clinical categories identified in the clinical file. This classification is exclusively based on the age deltas within the ages file. If the ages file comprises multiple systems, individual models are trained for each system, and a comprehensive model encompassing all system-specific age deltas is also trained.

over a wide age range of between 5 to 100 years there are many volumetric features that follow this u-shape [32], [33]. With feature extension volumetric features extended as x_i^2 or $x_i \times x_y$ become linearly correlated with age. Including the feature extension module in these cases could help to model these relationships in the age modeling task while using linear regression models instead of using more complex models such as SVM to capture this known interaction. Keeping a linear regression model with feature extension can allow for more explainability.

The pipeline is trained by minimizing the mean squared error between the predicted and chronological age. The pipeline undergoes training via cross-validation, with a default of 5-folds. The reported metrics are derived from these cross-validation folds. Users have the option to configure the hyperparameters of the model manually, or they can choose to do hyperparameter tuning. Hyperparameter tuning can be performed by choosing a specific model and optimizing the hyperparameters using cross validation and a grid search with a specified number of points. Alternatively, it can let AgeML choose the best model and preprocessing leveraging the capabilities of the hyperopt-sklearn library [34].

A well-known bias in age modeling is the tendency for younger participants to be assigned higher ages, while older participants are assigned lower ages than their actual ages. This phenomenon is attributed to a regression to the mean problem [21], [35], [36]. This bias can be conceptualized as our best estimate for a subject about whom we have no prior information being the mean age. The predicted age bias can be can be corrected [35], by taking into account the actual participant's age through fitting the following linear model :

$$\hat{Y} = \alpha \times Y + \beta. \tag{2}$$

$\hat{Y}$ is the predicted age. The coefficients α and β correspond to the slope and intercept of the linear model. The training set predictions $\hat{Y}_{train}$ and chronological age Y_{train} is used to fit these coefficients α and β. Selecting the appropriate method for age bias correction is of paramount importance to prevent the introduction of new biases, such as inflating correlation

effects. An alternative method to that used in this study is to use age as a covariate when studying group differences [35]. The fitted α and β are then used to correct the predictions in a test set to obtain an age bias corrected prediction $\tilde{Y}$ using:

$$\tilde{Y} = \hat{Y} + [Y - \alpha \times Y + \beta]. \tag{3}$$

Research in this context typically focused on the age gap or delta (δ), which is defined as the difference between the subject's predicted age after age bias correction and the subject's chronological age:

$$\delta = \tilde{Y} - Y. \tag{4}$$

AgeML provides the capability to simultaneously train distinct models for each system defined, thereby obtaining age deltas for each system. This is achieved by specifying which features belong to each system. Additionally, separate models can be trained based on a covariate, such as sex. This would train distinct models for males and females for each system.

Users also have the option to specify which participants are considered as healthy controls for model training purposes. The model is then trained exclusively on these designated participants and subsequently applied to calculate deltas for the remaining participants. The predicted age ($\hat{Y}$), the predicted age after age bias correction ($\tilde{Y}$) and the age delta (δ) are saved and can be used as inputs for further steps or for independent further analysis by the researcher, referred to as the *ages* file.

To ensure standardization and reproducibility, after each model is trained, both training and test metrics are stored. These metrics encompass Mean Absolute Error (MAE), Root Mean Square Error (RMSE), R^2, and Pearson's correlation (ρ). Recognizing that MAE and RMSE are strongly influenced by the age range of the training data, an additional metric referred to as the *dummy regressor* (DR) is calculated. The DR is obtained by using the mean age of control subjects as predictor for all subjects and calculating the resulting MAE and RMSE.

B. Factor Correlation

Previous research has addressed the relationship between the deltas and other external factors such as lifestyle choices or genetics. To visualize and quantify this relations, a Pearson correlation is calculated between each factor (Z) and each delta (δ) as well as the corresponding p-values. To account for multiple comparisons, p-values can be adjusted using either Bonferroni or False Discovery Rate (FDR) corrections. Additionally, users have the option to provide covariates (W) and remove covariate effects from the factors prior to conducting the correlation analysis. The covariate correction is calculated as the coefficients in the maximum likelihood solution of a general linear model, ie.,

$$\mathbf{\Gamma} = (\mathbf{W}^T\mathbf{W})^{-1}\mathbf{W}^T\mathbf{Z}. \tag{5}$$

As previously mentioned, $\mathbf{\Gamma}$ can be calculated based on only healthy controls, the whole population, or for each clinical group. Providing different covariate correction schemes enables accounting for complex covariate effects across diverse clinical populations, as their characteristics and their relation with covariates can vary significantly. The participant's covariate-corrected factors are then given by:

$$\hat{\mathbf{Z}} = \mathbf{Z} - \mathbf{W}\mathbf{\Gamma}. \tag{6}$$

The correlation analysis between factors and deltas is subsequently carried out using the corrected factors rather than the non-corrected ones.

C. Clinical Groups

AgeML offers the capability to examine the participant's age deltas labeled into various clinical populations. It assesses the distributions of age deltas for each provided clinical population, displaying the mean and standard deviation for each group. Additionally, AgeML determines whether the delta distributions exhibit significant differences between groups using a paired t-Test, indicating whether p-values are significant to Bonferroni or FDR corrections for multiple comparisons. Effect sizes, confidence intervals, and the delta mean differences are also reported by AgeML. Similar to the factor correlation analysis, covariate correction can be applied to the deltas previous to calculating distributions for each clinical group.

AgeML also integrates a quality control step to check whether the age distributions of participants within each clinical group exhibit significant differences as training on a specific distribution and subsequently making predictions in a different one with a different age can potentially introduce biases [21]. This is particularly evident when considering young and old populations, as many phenotypes follow quadratic trends, as observed in brain volumes [32], [33] and cognitive performance [37]. Hence, training on a young population and applying the model to an older one can be problematic and susceptible to bias.

D. Clinical Classification

In scenarios involving multiple organ-age systems, we represent $\mathbf{\Delta} \equiv \delta_i$ as the collection of deltas for each subject, and use them as the input for a classifier designed to categorize participants into various clinical groups. Employing a cross-validation strategy, a logistic regressor is trained to differentiate between two specific labels. The sigmoid function is then given by $\sigma(\mathbf{\Delta}) = 1/\left(1 + e^{-\beta_0 - \sum_i \beta_i \delta_i}\right)$, that after fitting the logistic regressor, the coefficient β_i accounts for the contribution of δ_i in the classification. The vector $\beta \equiv \beta_i$, previously normalized between 0 and 1 by dividing each component by the maximum β_i, is stored and printed to assess which organ delta participates with a higher relevance in the classification task.

V. Results

A. Reproducibility and Validation

To demonstrate that AgeML can replicate previous research findings, we chose to replicate the study by Tian et al. [1]. In this previous study a body age model is built along

TABLE II
MEAN ABSOLUTE ERROR (MAE) AND PEARSON'S CORRELATION (ρ) OF SVR MODELS FOR PREDICTING THE CHRONOLOGICAL AGES OF HEALTHY CONTROLS USING DIFFERENT PHYSIOLOGICAL SYSTEMS, COMPARING THE WORK OF TIAN ET AL. 2023 [1] WITH OURS.

METRIC	GENDER	MODEL	BODY	CARDIOVASCULAR	PULMONARY	MUSCULOSKELETAL
MAE	FEMALE	OURS	3.86 ± 0.21	5.38 ± 0.34	5.34 ± 0.27	4.98 ± 0.17
		TIAN ET AL. 2023 [1]	3.708 ± 0.0015	5.620 ± 0.0016	5.675 ± 0.0008	5.267 ± 0.0010
	MALE	OURS	4.61 ± 0.21	5.80 ± 0.25	5.55 ± 0.23	5.61 ± 0.22
		TIAN ET AL. 2023 [1]	4.463 ± 0.0016	6.323 ± 0.0011	6.130 ± 0.0008	6.052 ± 0.0012
ρ	FEMALE	OURS	0.753 ± 0.034	0.424 ± 0.065	0.443 ± 0.064	0.519 ± 0.045
		TIAN ET AL. 2023 [1]	0.787 ± 0.0002	0.436 ± 0.0004	0.421 ± 0.0003	0.513 ± 0.0002
	MALE	OURS	0.638 ± 0.033	0.296 ± 0.072	0.382 ± 0.051	0.372 ± 0.061
		TIAN ET AL. 2023 [1]	0.715 ± 0.0002	0.317 ± 0.0004	0.375 ± 0.0003	0.403 ± 0.0003

METRIC	GENDER	MODEL	IMMUNE	RENAL	HEPATIC	METABOLIC
MAE	FEMALE	OURS	5.70 ± 0.22	5.16 ± 0.24	5.13 ± 0.29	4.97 ± 0.24
		TIAN ET AL. 2023 [1]	5.965 ± 0.0019	5.220 ± 0.0012	5.466 ± 0.0010	5.302 ± 0.0022
	MALE	OURS	5.78 ± 0.22	5.41 ± 0.31	5.75 ± 0.25	5.85 ± 0.24
		TIAN ET AL. 2023 [1]	6.330 ± 0.0022	5.830 ± 0.0011	6.294 ± 0.0012	6.514 ± 0.0019
ρ	FEMALE	OURS	0.284 ± 0.087	0.475 ± 0.069	0.476 ± 0.084	0.524 ± 0.064
		TIAN ET AL. 2023 [1]	0.303 ± 0.0005	0.532 ± 0.0002	0.451 ± 0.0004	0.498 ± 0.0005
	MALE	OURS	0.297 ± 0.067	0.437 ± 0.058	0.308 ± 0.065	0.247 ± 0.065
		TIAN ET AL. 2023 [1]	0.301 ± 0.0007	0.453 ± 0.0003	0.322 ± 0.0004	0.230 ± 0.0009

* the reported $\pm$ are standard deviations of the MAE and Pearson's correlation across CV splits for our work and for Tian et. al. 2023 [1] they are the standard deviations of the MAE and Pearson's correlation over all subjects repeating the modeling 10 times with different seeds for the CV scheme.

with 7 separate organ system age models for the cardiovascular, pulmonary, musculoskeletal, immune, renal, hepatic and metabolic systems. The different age models are trained on healthy participants with no known chronic disease. The feature set consists of the same 78 body phenotypes described in Data, and participants with any missing phenotypes are excluded from the analysis. Training is carried out using a Support Vector Regressor (SVR) with a 20-fold cross-validation scheme (with a linear kernel, a regularization parameter of 1.0 and a gap tolerance of 0.001). Comparisons of the results between AgeML (Ours) and the one in Tian et al. 2023 [1] are given in Table II.

Tian et al. 2023 [1] cite the standard deviation of the Mean Absolute Error (MAE) and Pearson's correlation (ρ) over all subjects repeating the modeling 10 times with different seeds for the CV scheme. AgeML calculates the standard deviation of the MAE and ρ across the CV splits. This measure is larger but more representative of variability when the model looks at new data compared to the standard deviation of repeating the same experiment 10 times with the same data but different CV splits. It is also computationally less expensive. AgeML results were within 2 of our standard deviations measures of the results reported in Tian et al. 2023 [1]. There are some minor discrepancies possibly attributed to the differences in sample sizes (Ours N=6,388; Tian's N=28,589). Differences in sample size are due to differences in inclusion criteria, primarily our more stringent definition of chronic conditions.

We evaluated the performance of three different machine learning models (Ridge, Random Forest and Support Vector Machine) implemented in AgeML for age prediction across different organ systems. MAE and ρ for each organ system and model can be found in Supplementary File Table S.I. Across all organ systems, the error metrics for these models were generally within one standard deviation of each other, indicating similar performance. However, for the body model, which involved the most features, Ridge Regression demonstrated the best performance, while Random Forests exhibited the worst.

We have also used AgeML with data from ADNI [38] to reproduce our previous work Garcia Condado et al. 2023 [7]. We obtain similar error metrics in the brain age modeling task when using structural and neuropsychological features as seen in the Supplementary File Table S.II. We also obtain similar distributions of the deltas for the three different clinical groups: Controls (CN), Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD) as seen in the Supplementary File Table S.III.

B. Genetic Analysis

PRSs measure genetic risks for various diseases, and some diseases are closely related to specific organs, possibly influencing organ age. For example, it is expected that genetic predisposition to factors affecting the heart, such as cardiovascular diseases, coronary artery diseases, ischemic stroke, or hypertension, would lead to higher cardiovascular ages. Similarly, protective genetic factors that promote stronger bones with higher mineral density can lead to a reduction in musculoskeletal age. Here, and for the group of healthy participants, the correlation between age deltas for each organ and both Standard PRS and Enhanced PRS were compared, cf. Fig 2 shows statistically significant correlations between PRS and organ deltas (detailed correlation values and corresponding p-values are given in Supplementary File Table S.IV for Standard PRS and Table S.V for Enhanced PRS). Very remarkably, for the results corresponding to Enhanced PRS,

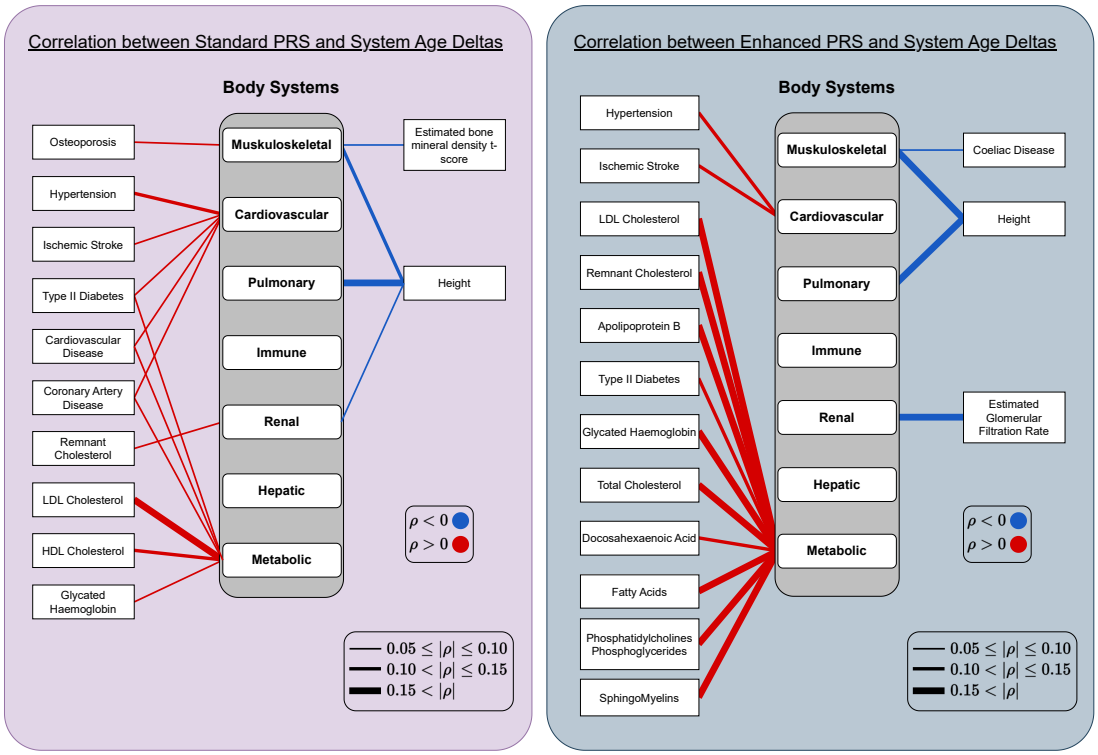

Fig. 2. **Factor correlation analysis between Polygenetic Risk Scores (PRS) and system-specific age deltas**. Each link represents a statistical significant Bonferroni-corrected Pearson's correlation ($p < 0.05$) and only those with $|\rho| > 0.05$ are plotted. The width of the link indicates the strength of the correlation. Red links denote positive correlations, which are associated with higher age deltas or accelerated aging. Blue links, on the other hand, represent negative correlations, suggesting protective or age-slowing effects. The left purple panel shows results for Standard PRSs (6,323 participants), while the right gray panel does it for Enhanced PRSs (1,234 participants). Note that several PRS labels are shared across panels, reassuring the significance of the correlations. All the Pearson's correlation values and their associated p-values can be found in the code repository.

we found that many metabolic genetic propensities negatively impact metabolic age, and in contrast, a protective effect of glomerular filtration. The differences in statistical significance between Standard PRS and Enhanced PRS may be due to the lower number of participants available for analysis in the latter.

C. Clinical data

Deltas for each organ systems were calculated for all participants. The average deltas for each clinical group and organ were calculated, and then compared across groups (Table III). The significance of these differences indicates that the participants belonging to clinical groups related to specific organs exhibit significant variations as compared to healthy controls. For example, individuals in the clinical category of circulatory diseases show a higher delta for the cardiovascular system compared to healthy controls. Detailed information about the differences in average deltas and their corresponding p-values, effect sizes, and confidence intervals between clinical groups and organ systems can be found in the Supplementary File Tables S.VI and S.VII, respectively.

Logistic regressor classifiers were trained to differentiate between healthy controls and each clinical category based on all organ deltas as input. Fig 3 displays the logistic regressor weights, which have been normalized for each case. Similar to the differences observed in average deltas for organ systems in specific clinical categories, the most discriminant features

TABLE III
DISEASE CATEGORIES THAT HAVE A STATISTICALLY SIGNIFICANT DIFFERENCE IN THEIR AGE DELTA COMPARED TO HEALTHY CONTROLS, FOR DIFFERENT SYSTEMS (BONFERRONI CORRECTED, $p < 0.05$).

SYSTEM DELTA	CLINICAL CATEGORY DIFFERENT TO CN				
CARDIOVASCULAR	CIR	MBD	MUS	RSP	NEO
PULMONARY	CIR	MBD	RSP	NEO	
METABOLIC	MUS	NEO			
IMMUNE	MBD				

CN: Healthty Controls, CIR: Circulatory system disease, MBD: neurodevelopmental disorders, MUS: Musculoskeletal system disorders, RSP: respiratory system diseases, NEO: neoplasms, BLD: blood diseases

between healthy controls and clinical categories related to specific organs are the deltas for those respective organs. Thus for instance, when classifying between healthy controls and cardiovascular disease, the most discriminant feature is the cardiovascular delta. To avoid biases in predictions, the majority class was under-sampled to match the size of the minority class.

VI. DISCUSSION

The first step towards standardization is showing the reproducibility of previous findings. In this context, we have replicated the study presented by Tian et al. [1]. AgeML facilitates standardization by providing a tool designed for use

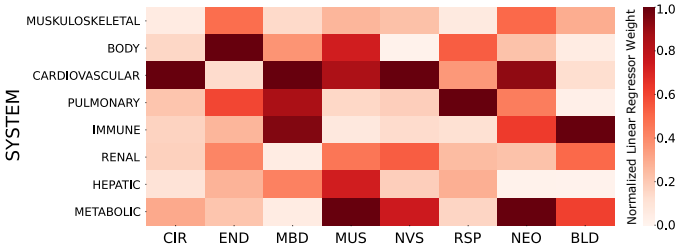

Fig. 3. **Normalized Logistic Regressor weights of system-specific deltas for classification between healthy controls versus a specific disease category**. The highest values for each column indicate the most relevant system-specific delta used by the Logistic Regressor for classifying healthy controls against the column disease category. Normalization is done per column (classifier), dividing the absolute values by the absolute highest one. Thus, the cells with the value 1.0 correspond to the absolute highest weight per column. In the situation of class imbalance, the majority class was undersampled to match the minority class occurrences.

with explicitly defined pipelines. Moreover, the reporting of standardized metrics by AgeML constitutes a crucial issue for usability. AgeML aims to establish the minimum guidelines for clear communication of training outcomes.

Previous research like BASE [22] focused on brain age modeling has shown the need for standardization and reproducibility. While BASE is tailored to brain age modeling, it highlights the significance of integrating standards into the field. AgeML not only replicates many of the benchmarks established in BASE with respect to metric reporting but also extends it further by enhancing ease of access, embracing open-source development, and demonstrating applicability across various types of clinical or organ data.

When exploring the performance of various machine learning models for age prediction, extending our analysis beyond Support Vector Regression (SVR) to include Ridge Regression and Random Forests, the results demonstrated that, across most organ systems, the error metrics for these models were within one standard deviation of each other. This suggests that the choice of model had minimal impact on predictive accuracy in these cases. In the body model, which featured the highest number of features, Ridge Regression outperformed the other models, while Random Forests performed the worst. This highlights the importance of considering the complexity and feature density of datasets when selecting a machine learning model. Although we applied hyperparameter tuning, performance improvements were minimal, consistent with findings from previous research [1]. These results suggest that while hyperparameter tuning and model selection remain important considerations, the inherent characteristics of the dataset may play a more significant role in determining model success. Future work should continue exploring model performance across a wider variety of datasets to further refine these predictions.

A potential limitation of AgeML is the risk of overfitting when constructing biological age models using only healthy cohorts and applying these models to both healthy and patient populations. This could affect the model's generalizability and the validity of comparisons between groups. It is an open discussion in the field whether models should be trained on a healthy population or a more diverse dataset representative of the general population. To address this, we propose two strategies. First, models can be trained on a broader population that includes both healthy and patient groups and is representative of the general population under study. Researchers must ensure this training set is separate from the test set. Second, AgeML incorporates regularization techniques and cross-validation to mitigate overfitting during model construction. Both approaches can be applied simultaneously for more robust model performance. While these measures help reduce overfitting, users should interpret results carefully when dealing with heterogeneous datasets. We acknowledge this limitation and suggest future improvements, including the integration of pre-trained models derived from larger, more diverse populations, which could enhance generalizability across various cohorts.

We have shown that genetic information, specifically Polygenic Risk Scores (PRS), are associated with distinct forms of organ aging, aligning with prior GWAS studies [9]. This is of critical importance for quantifying the extent to which an individual's organ age-variations can be attributed to genetics. Moreover, it paves the road for exploring therapeutic targets aimed at decelerating aging [9]. A logical continuation of this work is to develop organ-specific PRS based on genomic data to assess genetic predisposition to either significantly rejuvenated or minimally aged organs as a measure of aging resilience. Efforts in this direction are already in progress, notably in the context of brain age PRS [39].

The Standard PRS is derived from external datasets, ensuring broader applicability across populations but potentially sacrificing some degree of accuracy. In contrast, the Enhanced PRS incorporates additional training data from the UK Biobank itself, improving prediction accuracy but limiting its generalizability outside the UK Biobank population due to cohort-specific overfitting concerns. One practical implication of these differences lies in the sample size disparity: the Enhanced PRS (1,234 participants) may offer increased predictive power due to the enriched dataset, but the smaller sample size relative to the Standard PRS (6,323 participants) reduces its generalizability and increases the risk of bias. The smaller sample size of the Enhanced PRS could result in lower statistical significance, limiting the robustness of conclusions drawn from it. On the other hand, the Enhanced PRS may be less prone to biases from external datasets, offering a more tailored risk prediction in the UK Biobank cohort. To mitigate overfitting, we encourage the use of cross-validation and suggest that users carefully interpret results depending on the context in which these models are applied.

As expected the age deltas of each organ systems vary among clinical populations. This variation is particularly evident in the deltas associated with the cardiovascular system and its corresponding clinical category of cardiovascular disease. Similar to previous work [1], our results confirm that while large age deltas are not exclusive to any single disease, distinct patterns of organ deltas are associated with each specific disease.

The practical integration of age modeling in clinical environments is currently in a conceptual phase. An initial proposition involves using age deltas as a flag, where during

routine analyses, these deltas are computed, and an alert is triggered if any of the calculated deltas are high (previously determined a normal or standardized distribution of deltas). Although organ deltas lack specificity to particular diseases (eg., patients with Parkinson's, Alzheimer's, or depression might exhibit an elevated brain delta), they can alert clinicians to potential deviations from healthy aging normative patterns. By computing deltas for various organs and achieving a multimodal perspective, it is also feasible to mitigate the issue of non-specificity, thereby enhancing the feasibility of age deltas in clinical settings.

AgeML is a first approach to setting a standard in the community and a minimal software tool for basic age modeling. The AgeML project will continue by integrating more features. Currently AgeML supports only tabular data; however, future plans aim to incorporate additional data modalities, including imaging data (e.g., MRI or CT scans) routinely used in medical settings, as well as time series data, such as electrocardiograms (ECG). The next step involves integrating foundational models that have been trained on extensive datasets, such as the UKBiobank. Following that, the strategy includes providing transfer learning capabilities to retrain models with local datasets. In medical contexts, the demand for interpretability is paramount and should be addressed by using for example Shap values [40]. The factor correlation pipeline could be upgraded to introduce causal analysis. Ultimately, it is anticipated that community-driven efforts will emerge, aiming to incorporate a broader range of desired features and improvements.

VII. Conclusion

We have shown that AgeML is capable to replicate prior studies in the field, thereby paving the way for standardization of age modeling strategies. We have proven the AgeML efficacy by investigating the interplay between age-related organ variations and genetics, and in particular, by assessing the correlation between Polygenic Risk Scores and specific organ deltas. Consistent with previous work, statistical significant differences in organ deltas occur between healthy controls and clinical populations, especially when focusing on organ-systems related to specific diseases.

The ultimate objective of our study extends beyond merely developing a tool; it aims to trigger a community of developers who contribute new features and help lead towards age-modeling standardization. The major goal of AgeML is to standardize procedures, reduce the barriers to entry for age modeling, and guarantee the reproducibility of research findings. The AgeML project is Open-Source, designed to foster a welcoming environment and a community to work together to improve and validate existing methodologies. We are actively inviting new developers interested in contributing to the growth and expansion of the package.

Acknolwedgments

This research has been conducted using the UKBiobank Resource under Application Number 79105.

References

[1] Y. E. Tian, V. Cropley, A. B. Maier, N. T. Lautenschlager, M. Breakspear, and A. Zalesky, "Heterogeneous aging across multiple organ systems and prediction of chronic disease and mortality," *Nature Medicine*, Apr. 2023. [Online]. Available: https://www.nature.com/articles/s41591-023-02296-6

[2] K. Franke, G. Ziegler, S. Klöppel, and C. Gaser, "Estimating the age of healthy subjects from T1-weighted MRI scans using kernel methods: Exploring the influence of various parameters," *NeuroImage*, vol. 50, no. 3, pp. 883–892, Apr. 2010. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811910000108

[3] P. Bonifazi, A. Erramuzpe, I. Diez, I. Gabilondo, M. P. Boisgontier, L. Pauwels, S. Stramaglia, S. P. Swinnen, and J. M. Cortes, "Structure–function multi-scale connectomics reveals a major role of the fronto-striato-thalamic circuit in brain aging," *Human Brain Mapping*, vol. 39, no. 12, pp. 4663–4677, 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.24312

[4] A. Erramuzpe, R. Schurr, J. D. Yeatman, I. H. Gotlib, M. D. Sacchet, K. E. Travis, H. M. Feldman, and A. A. Mezer, "A Comparison of Quantitative R1 and Cortical Thickness in Identifying Age, Lifespan Dynamics, and Disease States of the Human Cortex," *Cerebral Cortex (New York, N.Y.: 1991)*, vol. 31, no. 2, pp. 1211–1226, Jan. 2021.

[5] A. O. Ladejobi, J. R. Medina-Inojosa, M. Shelly Cohen, Z. I. Attia, C. G. Scott, N. K. LeBrasseur, B. J. Gersh, P. A. Noseworthy, P. A. Friedman, S. Kapa, and F. Lopez-Jimenez, "The 12-lead electrocardiogram as a biomarker of biological age," *European Heart Journal - Digital Health*, vol. 2, no. 3, pp. 379–389, Sep. 2021. [Online]. Available: https://academic.oup.com/ehjdh/article/2/3/379/6248088

[6] J. Ma, X. Xu, M. Li, Y. Zhang, L. Zhang, P. Ma, J. Hou, Y. Lei, J. Liu, X. Huangfu, Y. Yang, X. Yi, G. Cheng, J. Bai, X. Zhong, X. Xu, and Y. Wang, "Predictive models of aging of the human eye based on ocular anterior segment morphology," *Journal of Biomedical Informatics*, vol. 120, p. 103855, Aug. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1532046421001842

[7] J. Garcia Condado, J. M. Cortes, and for the Alzheimer's Disease Neuroimaging Initiative, "NeuropsychBrainAge: A biomarker for conversion from mild cognitive impairment to Alzheimer's disease," *Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring*, vol. 15, no. 4, p. e12493, Oct. 2023. [Online]. Available: https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/dad2.12493

[8] C. Sudlow, J. Gallacher, N. Allen, V. Beral, P. Burton, J. Danesh, P. Downey, P. Elliott, J. Green, M. Landray, B. Liu, P. Matthews, G. Ong, J. Pell, A. Silman, A. Young, T. Sprosen, T. Peakman, and R. Collins, "UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age," *PLOS Medicine*, vol. 12, no. 3, p. e1001779, Mar. 2015. [Online]. Available: https://dx.plos.org/10.1371/journal.pmed.1001779

[9] J. Wen, Y. E. Tian, I. Skampardoni, Z. Yang, E. Mamourian, F. Anagnostakis, B. Zhao, A. W. Toga, A. Zaleskey, and C. Davatzikos, "The Genetic Architecture of Biological Age in Nine Human Organ Systems," Genetic and Genomic Medicine, preprint, Jun. 2023. [Online]. Available: http://medrxiv.org/lookup/doi/10.1101/2023.06.08.23291168

[10] A. Bashan, R. P. Bartsch, J. W. Kantelhardt, S. Havlin, and P. C. Ivanov, "Network physiology reveals relations between network topology and physiological function," *Nature Communications*, vol. 3, no. 1, p. 702, Feb. 2012. [Online]. Available: https://www.nature.com/articles/ncomms1705

[11] C. Gaser, K. Franke, S. Klöppel, N. Koutsouleris, H. Sauer, and Alzheimer's Disease Neuroimaging Initiative, "BrainAGE in Mild Cognitive Impaired Patients: Predicting the Conversion to Alzheimer's Disease," *PLoS ONE*, vol. 8, no. 6, p. e67346, Jun. 2013. [Online]. Available: https://dx.plos.org/10.1371/journal.pone.0067346

[12] G. Hwang, A. Abdulkadir, G. Erus, M. Habes, R. Pomponio, H. Shou, J. Doshi, E. Mamourian, T. Rashid, M. Bilgel, Y. Fan, A. Sotiras, D. Srinivasan, J. C. Morris, M. S. Albert, N. R. Bryan, S. M. Resnick, I. M. Nasrallah, C. Davatzikos, D. A. Wolk, from the iSTAGING consortium, and for the ADNI, "Disentangling Alzheimer's disease neurodegeneration from typical brain ageing using machine learning," *Brain Communications*, vol. 4, no. 3, p. fcac117, May 2022. [Online]. Available: https://academic.oup.com/braincomms/article/doi/10.1093/braincomms/fcac117/6582261

[13] A. Pichet Binette, J. Gonneaud, J. W. Vogel, R. La Joie, P. Rosa-Neto, D. L. Collins, J. Poirier, J. C. S. Breitner, S. Villeneuve, E. Vachon-Presseau, for the Alzheimer's Disease Neuroimaging Initiative, and the PREVENT-AD Research Group, "Morphometric network differences in ageing versus Alzheimer's disease dementia,"

Brain, vol. 143, no. 2, pp. 635–649, Feb. 2020. [Online]. Available: https://academic.oup.com/brain/article/143/2/635/5732992

[14] D. Vidal-Pineiro, Y. Wang, S. K. Krogsrud, I. K. Amlien, W. F. Baaré, D. Bartres-Faz, L. Bertram, A. M. Brandmaier, C. A. Drevon, S. Düzel, K. Ebmeier, R. N. Henson, C. Junqué, R. A. Kievit, S. Kuhn, E. Leonardsen, U. Lindenberger, K. S. Madsen, F. Magnussen, A. M. Mowinckel, L. Nyberg, J. M. Roe, B. Segura, S. M. Smith, O. Sorensen, S. Suri, R. Westerhausen, A. Zalesky, E. Zsoldos, K. B. Walhovd, and A. Fjell, "Individual variations in 'brain age' relate to early-life factors more than to longitudinal brain change," *eLife*, vol. 10, p. e69995, Nov. 2021. [Online]. Available: https://elifesciences.org/articles/69995

[15] V. M. Bashyam, G. Erus, J. Doshi, M. Habes, I. M. Nasrallah, M. Truelove-Hill, D. Srinivasan, L. Mamourian, R. Pomponio, Y. Fan, L. J. Launer, C. L. Masters, P. Maruff, C. Zhuo, H. Völzke, S. C. Johnson, J. Fripp, N. Koutsouleris, T. D. Satterthwaite, D. Wolf, R. E. Gur, R. C. Gur, J. Morris, M. S. Albert, H. J. Grabe, S. Resnick, R. N. Bryan, D. A. Wolk, H. Shou, and C. Davatzikos, "MRI signatures of brain age and disease over the lifespan based on a deep brain network and 14 468 individuals worldwide," *Brain*, vol. 143, no. 7, pp. 2312–2324, Jul. 2020. [Online]. Available: https://academic.oup.com/brain/article/143/7/2312/5863667

[16] T. Hahn, L. Fisch, J. Ernsting, N. R. Winter, R. Leenings, K. Sarink, D. Emden, T. Kircher, K. Berger, and U. Dannlowski, "From 'loose fitting' to high-performance, uncertainty-aware brain-age modelling," *Brain*, vol. 144, no. 3, pp. e31–e31, Apr. 2021. [Online]. Available: https://academic.oup.com/brain/article/144/3/e31/6214475

[17] S. More, G. Antonopoulos, F. Hoffstaedter, J. Caspers, S. B. Eickhoff, and K. R. Patil, "Brain-age prediction: A systematic comparison of machine learning workflows," *NeuroImage*, vol. 270, p. 119947, Apr. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811923000940

[18] L. Baecker, R. Garcia-Dias, S. Vieira, C. Scarpazza, and A. Mechelli, "Machine learning for brain age prediction: Introduction to methods and clinical applications," *eBioMedicine*, vol. 72, p. 103600, Oct. 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2352396421003935

[19] A. Puc, V. Struc, and K. Grm, "Analysis of Race and Gender Bias in Deep Age Estimation Models," in *2020 28th European Signal Processing Conference (EUSIPCO)*. Amsterdam, Netherlands: IEEE, Jan. 2021, pp. 830–834. [Online]. Available: https://ieeexplore.ieee.org/document/9287219/

[20] S. Moguilner, S. Baez, H. Hernandez, J. Migeot, A. Legaz, R. Gonzalez-Gomez, F. R. Farina, P. Prado, J. Cuadros, E. Tagliazucchi, F. Altschuler, M. A. Maito, M. E. Godoy, J. Cruzat, P. A. Valdes-Sosa, F. Lopera, J. F. Ochoa-Gómez, A. G. Hernandez, J. Bonilla-Santos, R. A. Gonzalez-Montealegre, R. Anghinah, L. E. d'Almeida Manfrinati, S. Fittipaldi, V. Medel, D. Olivares, G. G. Yener, J. Escudero, C. Babiloni, R. Whelan, B. Güntekin, H. Yırıkoğulları, H. Santamaria-Garcia, A. F. Lucas, D. Huepe, G. Di Caterina, M. Soto-Añari, A. Birba, A. Sainz-Ballesteros, C. Coronel-Oliveros, A. Yigezu, E. Herrera, D. Abasolo, K. Kilborn, N. Rubido, R. A. Clark, R. Herzog, D. Yerlikaya, K. Hu, M. A. Parra, P. Reyes, A. M. García, D. L. Matallana, J. A. Avila-Funes, A. Slachevsky, M. I. Behrens, N. Custodio, J. F. Cardona, P. Barttfeld, I. L. Brusco, M. A. Bruno, A. L. Sosa Ortiz, S. D. Pina-Escudero, L. T. Takada, E. Resende, K. L. Possin, M. O. De Oliveira, A. Lopez-Valdes, B. Lawlor, I. H. Robertson, K. S. Kosik, C. Duran-Aniotz, V. Valcour, J. S. Yokoyama, B. Miller, and A. Ibanez, "Brain clocks capture diversity and disparities in aging and dementia across geographically diverse populations," *Nature Medicine*, Aug. 2024. [Online]. Available: https://www.nature.com/articles/s41591-024-03209-x

[21] A. G. de Lange, M. Anatürk, J. Rokicki, L. K. M. Han, K. Franke, D. Alnæs, K. P. Ebmeier, B. Draganski, T. Kaufmann, L. T. Westlye, T. Hahn, and J. H. Cole, "Mind the gap: Performance metric evaluation in brain-age prediction," *Human Brain Mapping*, vol. 43, no. 10, pp. 3113–3129, Jul. 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/hbm.25837

[22] L. Dular and Z. Spiclin, "BASE: Brain Age Standardized Evaluation," *NeuroImage*, vol. 285, p. 120469, Jan. 2024. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811923006183

[23] R. K. Gebre, M. L. Senjem, S. Raghavan, C. G. Schwarz, J. L. Gunter, E. I. Hofrenning, R. I. Reid, K. Kantarci, J. Graff-Radford, D. S. Knopman, R. C. Petersen, C. R. Jack, and P. Vemuri, "Cross–scanner harmonization methods for structural MRI may need further work: A comparison study," *NeuroImage*, vol. 269, p. 119912, Apr. 2023. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811923000605

[24] R. Pomponio, G. Erus, M. Habes, J. Doshi, D. Srinivasan, E. Mamourian, V. Bashyam, I. M. Nasrallah, T. D. Satterthwaite, Y. Fan, L. J. Launer, C. L. Masters, P. Maruff, C. Zhuo, H. Völzke, S. C. Johnson, J. Fripp, N. Koutsouleris, D. H. Wolf, R. Gur, R. Gur, J. Morris, M. S. Albert, H. J. Grabe, S. M. Resnick, R. N. Bryan, D. A. Wolk, R. T. Shinohara, H. Shou, and C. Davatzikos, "Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan," *NeuroImage*, vol. 208, p. 116450, Mar. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811919310419

[25] J.-P. Fortin, E. M. Sweeney, J. Muschelli, C. M. Crainiceanu, and R. T. Shinohara, "Removing inter-subject technical variability in magnetic resonance imaging studies," *NeuroImage*, vol. 132, pp. 198–212, May 2016. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1053811916001452

[26] HCUP. (2020) Hcup clinical classifications software refined (ccsr) for icd-10-cm diagnoses, v2021.2. [Online]. Available: www.hcup-us.ahrq.gov/toolssoftware/ccsr/dxccsr.jsp

[27] HCUP. (2021) Hcup chronic condition indicator for icd-10-cm (beta version), v2021.1. [Online]. Available: www.hcup-us.ahrq.gov/toolssoftware/chronic_icd10/chronic_icd10.jsp

[28] D. J. Thompson, D. Wells, S. Selzam, I. Peneva, R. Moore, K. Sharp, W. A. Tarran, E. J. Beard, F. Riveros-Mckay, C. Giner-Delgado, D. Palmer, P. Seth, J. Harrison, M. Futema, Genomics England Research Consortium, G. McVean, V. Plagnol, P. Donnelly, and M. E. Weale, "UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits," Genetic and Genomic Medicine, preprint, Jun. 2022. [Online]. Available: http://medrxiv.org/lookup/doi/10.1101/2022.06.16.22276246

[29] PyPI. (2021) Python package index - pypi. [Online]. Available: https://pypi.org/

[30] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and Édouard Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. 85, pp. 2825–2830, 2011. [Online]. Available: http://jmlr.org/papers/v12/pedregosa11a.html

[31] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco California USA: ACM, Aug. 2016, pp. 785–794. [Online]. Available: https://dl.acm.org/doi/10.1145/2939672.2939785

[32] D. Dima, A. Modabbernia, E. Papachristou, G. E. Doucet, I. Agartz, M. Aghajani, T. N. Akudjedu, A. Albajes-Eizagirre, D. Alnæs, K. I. Alpert, M. Andersson, N. C. Andreasen, O. A. Andreassen, P. Asherson, T. Banaschewski, N. Bargallo, S. Baumeister, R. Baur-Streubel, A. Bertolino, A. Bonvino, D. I. Boomsma, S. Borgwardt, J. Bourque, D. Brandeis, A. Breier, H. Brodaty, R. M. Brouwer, J. K. Buitelaar, G. F. Busatto, R. L. Buckner, V. Calhoun, E. J. Canales-Rodríguez, D. M. Cannon, X. Caseras, F. X. Castellanos, S. Cervenka, T. M. Chaim-Avancini, C. R. K. Ching, V. Chubar, V. P. Clark, P. Conrod, A. Conzelmann, B. Crespo-Facorro, F. Crivello, E. A. Crone, U. Dannlowski, A. M. Dale, C. Davey, E. J. C. Geus, L. Haan, G. I. Zubicaray, A. Braber, E. W. Dickie, A. Di Giorgio, N. T. Doan, E. S. Dørum, S. Ehrlich, S. Erk, T. Espeseth, H. Fatouros-Bergman, S. E. Fisher, J. Fouche, B. Franke, T. Frodl, P. Fuentes-Claramonte, D. C. Glahn, I. H. Gotlib, H. Grabe, O. Grimm, N. A. Groenewold, D. Grotegerd, O. Gruber, P. Gruner, R. E. Gur, R. C. Gur, T. Hahn, B. J. Harrison, C. A. Hartman, S. N. Hatton, A. Heinz, D. J. Heslenfeld, D. P. Hibar, I. B. Hickie, B. Ho, P. J. Hoekstra, S. Hohmann, A. J. Holmes, M. Hoogman, N. Hosten, F. M. Howells, H. E. Hulshoff Pol, C. Huyser, N. Jahanshad, A. James, T. L. Jernigan, J. Jiang, E. G. Jönsson, J. A. Joska, R. Kahn, A. Kalnin, R. Kanai, M. Klein, T. P. Klyushnik, L. Koenders, S. Koops, B. Krämer, J. Kuntsi, J. Lagopoulos, L. Lázaro, I. Lebedeva, W. H. Lee, K. Lesch, C. Lochner, M. W. J. Machielsen, S. Maingault, N. G. Martin, I. Martínez-Zalacaín, D. Mataix-Cols, B. Mazoyer, C. McDonald, B. C. McDonald, A. M. McIntosh, K. L. McMahon, G. McPhilemy, S. Meinert, J. M. Menchón, S. E. Medland, A. Meyer-Lindenberg, J. Naaijen, P. Najt, T. Nakao, J. E. Nordvik, L. Nyberg, J. Oosterlaan, V. O. Foz, Y. Paloyelis, P. Pauli, G. Pergola, E. Pomarol-Clotet, M. J. Portella, S. G. Potkin, J. Radua, A. Reif, D. A. Rinker, J. L. Roffman, P. G. P. Rosa, M. D. Sacchet, P. S. Sachdev, R. Salvador, P. Sánchez-Juan, S. Sarró, T. D. Satterthwaite, A. J. Saykin, M. H. Serpa, L. Schmaal, K. Schnell, G. Schumann, K. Sim, J. W. Smoller, I. Sommer, C. Soriano-Mas, D. J. Stein, L. T. Strike, S. C. Swagerman, C. K. Tamnes, H. S. Temmingh, S. I. Thomopoulos, A. S. Tomyshev, D. Tordesillas-Gutiérrez, J. N. Trollor, J. A. Turner, A. Uhlmann, O. A. Heuvel, D. Meer, N. J. A.

Wee, N. E. M. Haren, D. Ent, T. G. M. Erp, I. M. Veer, D. J. Veltman, A. Voineskos, H. Völzke, H. Walter, E. Walton, L. Wang, Y. Wang, T. H. Wassink, B. Weber, W. Wen, J. D. West, L. T. Westlye, H. Whalley, L. M. Wierenga, S. C. R. Williams, K. Wittfeld, D. H. Wolf, A. Worker, M. J. Wright, K. Yang, Y. Yoncheva, M. V. Zanetti, G. C. Ziegler, P. M. Thompson, S. Frangou, and Karolinska Schizophrenia Project (KaSP), "Subcortical volumes across the lifespan: Data from 18,605 healthy individuals aged 3–90 years," *Human Brain Mapping*, vol. 43, no. 1, pp. 452–469, Jan. 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/hbm.25320

[33] R. A. I. Bethlehem, J. Seidlitz, S. R. White, J. W. Vogel, K. M. Anderson, C. Adamson, S. Adler, G. S. Alexopoulos, E. Anagnostou, A. Areces-Gonzalez, D. E. Astle, B. Auyeung, M. Ayub, J. Bae, G. Ball, S. Baron-Cohen, R. Beare, S. A. Bedford, V. Benegal, F. Beyer, J. Blangero, M. Blesa Cábez, J. P. Boardman, M. Borzage, J. F. Bosch-Bayard, N. Bourke, V. D. Calhoun, M. M. Chakravarty, C. Chen, C. Chertavian, G. Chetelat, Y. S. Chong, J. H. Cole, A. Corvin, M. Costantino, E. Courchesne, F. Crivello, V. L. Cropley, J. Crosbie, N. Crossley, M. Delarue, R. Delorme, S. Desrivieres, G. A. Devenyi, M. A. Di Biase, R. Dolan, K. A. Donald, G. Donohoe, K. Dunlop, A. D. Edwards, J. T. Elison, C. T. Ellis, J. A. Elman, L. Eyler, D. A. Fair, E. Feczko, P. C. Fletcher, P. Fonagy, C. E. Franz, L. Galan-Garcia, A. Gholipour, J. Giedd, J. H. Gilmore, D. C. Glahn, I. M. Goodyer, P. E. Grant, N. A. Groenewold, F. M. Gunning, R. E. Gur, R. C. Gur, C. F. Hammill, O. Hansson, T. Hedden, A. Heinz, R. N. Henson, K. Heuer, J. Hoare, B. Holla, A. J. Holmes, R. Holt, H. Huang, K. Im, J. Ipser, C. R. Jack, A. P. Jackowski, T. Jia, K. A. Johnson, P. B. Jones, D. T. Jones, R. S. Kahn, H. Karlsson, L. Karlsson, R. Kawashima, E. A. Kelley, S. Kern, K. W. Kim, M. G. Kitzbichler, W. S. Kremen, F. Lalonde, B. Landeau, S. Lee, J. Lerch, J. D. Lewis, J. Li, W. Liao, C. Liston, M. V. Lombardo, J. Lv, C. Lynch, T. T. Mallard, M. Marcelis, R. D. Markello, S. R. Mathias, B. Mazoyer, P. McGuire, M. J. Meaney, A. Mechelli, N. Medic, B. Misic, S. E. Morgan, D. Mothersill, J. Nigg, M. Q. W. Ong, C. Ortinau, R. Ossenkoppele, M. Ouyang, L. Palaniyappan, L. Paly, P. M. Pan, C. Pantelis, M. M. Park, T. Paus, Z. Pausova, D. Paz-Linares, A. Pichet Binette, K. Pierce, X. Qian, J. Qiu, A. Qiu, A. Raznahan, T. Rittman, A. Rodrigue, C. K. Rollins, R. Romero-Garcia, L. Ronan, M. D. Rosenberg, D. H. Rowitch, G. A. Salum, T. D. Satterthwaite, H. L. Schaare, R. J. Schachar, A. P. Schultz, G. Schumann, M. Schöll, D. Sharp, R. T. Shinohara, I. Skoog, C. D. Smyser, R. A. Sperling, D. J. Stein, A. Stolicyn, J. Suckling, G. Sullivan, Y. Taki, B. Thyreau, R. Toro, N. Traut, K. A. Tsvetanov, N. B. Turk-Browne, J. J. Tuulari, C. Tzourio, Vachon-Presseau, M. J. Valdes-Sosa, P. A. Valdes-Sosa, S. L. Valk, T. Van Amelsvoort, S. N. Vandekar, L. Vasung, L. W. Victoria, S. Villeneuve, A. Villringer, P. E. Vértes, K. Wagstyl, Y. S. Wang, S. K. Warfield, V. Warrier, E. Westman, M. L. Westwater, H. C. Whalley, A. V. Witte, N. Yang, B. Yeo, H. Yun, A. Zalesky, H. J. Zar, A. Zettergren, J. H. Zhou, H. Ziauddeen, A. Zugman, X. N. Zuo, 3R-BRAIN, AIBL, C. Rowe, Alzheimer's Disease Neuroimaging Initiative, Alzheimer's Disease Repository Without Borders Investigators, G. B. Frisoni, CALM Team, Cam-CAN, CCNP, COBRE, cVEDA, ENIGMA Developmental Brain Age Working Group, Developing Human Connectome Project, FinnBrain, Harvard Aging Brain Study, IMAGEN, KNE96, The Mayo Clinic Study of Aging, NSPN, POND, The PREVENT-AD Research Group, A. P. Binette, VETSA, E. T. Bullmore, and A. F. Alexander-Bloch, "Brain charts for the human lifespan," *Nature*, vol. 604, no. 7906, pp. 525–533, Apr. 2022. [Online]. Available: https://www.nature.com/articles/s41586-022-04554-y

[34] B. Komer, J. Bergstra, and C. Eliasmith, "Hyperopt-Sklearn: Automatic Hyperparameter Configuration for Scikit-Learn," in *Python in Science Conference*, Austin, Texas, 2014, pp. 32–37.

[35] A.-M. G. de Lange and J. H. Cole, "Commentary: Correction procedures in brain-age prediction," *NeuroImage: Clinical*, vol. 26, p. 102229, 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2213158220300668

[36] E. R. Butler, A. Chen, R. Ramadan, T. T. Le, K. Ruparel, T. M. Moore, T. D. Satterthwaite, F. Zhang, H. Shou, R. C. Gur, T. E. Nichols, and R. T. Shinohara, "Pitfalls in brain age analyses," *Human Brain Mapping*, vol. 42, no. 13, pp. 4092–4101, Sep. 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/hbm.25533

[37] C. Grady, "The cognitive neuroscience of ageing," *Nature Reviews Neuroscience*, vol. 13, no. 7, pp. 491–505, Jul. 2012. [Online]. Available: https://www.nature.com/articles/nrn3256

[38] R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, J. Q. Trojanowski, and M. W. Weiner, "Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical characterization," *Neurology*, vol. 74, no. 3, pp. 201–209, Jan. 2010. [Online]. Available: https://www.neurology.org/lookup/doi/10.1212/WNL.0b013e3181cb3e25

[39] J. Wen, B. Zhao, Z. Yang, G. Erus, I. Skampardoni, E. Mamourian, Y. Cui, G. Hwang, J. Bao, A. Boquet-Pujadas, Z. Zhou, Y. Veturi, M. D. Ritchie, H. Shou, P. M. Thompson, L. Shen, A. W. Toga, and C. Davatzikos, "The Genetic Architecture of Multimodal Human Brain Age," Genetics, preprint, Apr. 2023. [Online]. Available: http://biorxiv.org/lookup/doi/10.1101/2023.04.13.536818

[40] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS'17. Red Hook, NY, USA: Curran Associates Inc., 2017, p. 4768–4777.