Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ARTICLE TYPE

Uncertainty quantification in epigenetic clocks via conformalized quantile regression

Yanping Li[1] | **Jaclyn M. Goodrich**[2] | **Karen E Peterson**[3] | **Peter X-K Song**[4] | **Lan Luo**[5]

[1]School of Statistics and Data Science, Nankai University, China

[2]Department of Environmental Health Sciences, University of Michigan, Ann Arbor, USA

[3]Department of Nutritional Sciences, University of Michigan, Ann Arbor, USA

[4]Department of Biostatistics, University of Michigan, Ann Arbor, USA

[5]Department of Biostatistics and Epidemiology, Rutgers University, USA

Correspondence
Lan Luo, Department of Biostatistics and Epidemiology, Rutgers University.
Email: ll1118@sph.rutgers.edu

Funding Information
This research was supported by the National Institute on Aging of the National Institutes of Health (R21AG083364).

Abstract

DNA methylation (DNAm) is a chemical modification of DNA that can be influenced by various factors, including age, environment, and lifestyle. An epigenetic clock is a predictive tool that measures biological age based on DNAm levels. It can provide insights into an individual's biological age, which may differ from their chronological age. This difference, known as the epigenetic age acceleration, may indicate the state of one's health and risk for age-related diseases. Moreover, epigenetic clocks are used in studies of aging to assess the effectiveness of anti-aging interventions and to understand the underlying mechanisms of aging and disease. Various epigenetic clocks have been developed using samples from different populations, tissues, and cell types, typically by training high-dimensional linear regression models with an elastic net penalty. While these models can predict mean biological age with high precision, there is a lack of uncertainty quantification which is important for interpreting the precision of age estimations and for clinical decision-making. To understand the distribution of a biological age clock beyond its mean, we propose a general pipeline for training epigenetic clocks, based on an integration of high-dimensional quantile regression and conformal prediction, to effectively reveal population heterogeneity and construct prediction intervals. Our approach produces adaptive prediction intervals not only achieving nominal coverage but also accounting for the inherent variability across individuals. By using the data collected from 728 blood samples in 11 DNAm datasets from children, we find that our quantile regression-based prediction intervals are narrower than those derived from conventional mean regression-based epigenetic clocks. This observation demonstrates an improved statistical efficiency over the existing pipeline for training epigenetic clocks. In addition, the resulting intervals have a synchronized varying pattern to age acceleration, effectively revealing cellular evolutionary heterogeneity in age patterns in different developmental stages during individual childhoods and adolescent cohort. Our findings suggest that conformalized high-dimensional quantile regression can produce valid prediction intervals and uncover underlying population heterogeneity. Although our methodology focuses on the distribution of aging in children, it is applicable to a broader range of populations to improve understanding of epigenetic age beyond the mean. This inference-based toolbox could provide valuable insights for future applications of epigenetic interventions for age-related diseases.

KEYWORDS
biological age, conformal prediction, DNA methylation, epigenetic clock, heterogeneity, pediatrics

1 | BACKGROUND

While chronological age is arguably a strong risk factor for aging-related death and diseases, individuals of the same chronological age may exhibit great heterogeneity in physiologic functions and rate of biological aging. Identifying aging biomarkers is a crucial step in the evaluation of interventions aimed at promoting healthier aging. Epigenetic age is a biomarker of aging that has been reported to be associated with age-related disease and all-cause mortality [1,2,3,4]. It has been found that composite measures of DNA methylation (DNAm) levels across specific sets of cytosine-phosphate-guanine (CpG) sites, often called epigenetic clocks, are strongly associated with chronological age or age-related diseases. One of the first and most widely used

Abbreviations: DNAm, DNA methylation; CpG, cytosine-phosphate-guanine; GEO, Gene Expression Omnibus.

epigenetic age predictors is Horvath's epigenetic clock[1], a statistical prediction model that uses DNAm at 353 CpG sites to predict chronological age.

Standard approaches for training epigenetic clocks involve several key steps: (i) collecting biological samples from individuals with diverse backgrounds; (ii) extracting DNA and performing DNA methylation analysis; (iii) conducting data preprocessing procedures such as missing data imputation, outliers removal, and data normalization; (iv) adopting a feature screening method for identifying relevant CpG sites that are predictive of age or linked with aging processes; (v) fitting a high-dimensional regression model with elastic net penalty; and (vi) evaluating model performance on an independent test dataset to verify its accuracy and robustness.

Despite the well-established pipelines for constructing epigenetic clocks, most of them provide point mean predictions only[1,2,5]. In biological age prediction, it is important not only to predict accurately but also to quantify the uncertainty of the predictions. This is especially true in biomedical applications involving high-stakes decision making, such as predicting the risk of a disease. The uncertainty in a prediction can be quantified using a prediction interval, giving lower and upper bounds between which the response variable lies with high probability. More specifically, epigenetic clocks measure biological age, which can differ significantly from chronological age due to various factors such as lifestyle, genetics, and environment. Prediction intervals help in interpreting these predictions by indicating how much variation there is around the estimated biological age, thus aiding in better understanding the potential impacts of these factors. In addition, when clinicians or researchers use epigenetic clocks to study aging or assess the risk of age-related diseases, prediction intervals can help them assess the reliability of the clock in different populations or under different conditions. A narrow interval suggests high reliability and vice versa, influencing the confidence in using these tools for medical or research decisions. However, it is worth noting that existing epigenetic clocks are primarily built from ultra high-dimensional DNAm data where most inferential methods rely on nontrivial assumptions such as the linear model being true, the error distribution, the homoscedasticity of errors and so on[6,7,8,9].

In high-dimensional settings, where the number of predictors (CpGs) exceeds the number of subjects, traditional statistical methods often struggle to provide reliable uncertainty estimates due to overfitting and high variance. Conformal prediction methods can mitigate these issues by constructing prediction intervals that are valid under a minimal set of assumptions about the data distribution[10]. They can be applied on top of any predictive model, and this flexibility is particularly useful in constructing prediction intervals for epigenetic clocks. A key component of conformal prediction is the nonconformity score, which quantifies how well a new data point conforms to previously observed data. In the field of epigenetic aging, it can be the residual from regressing an epigenetic clock on chronological age, referred to as epigenetic age acceleration which occurs when an individual's DNAm age exceeds their chronological age. Prior studies suggest that older epigenetic age may be associated with lower levels of physical functioning, and declines in global cognitive functioning among long-lived individuals[3,4,11,12,13]. Since epigenetic age acceleration has been found to be closely related to various adverse outcomes such as cardiovascular diseases and cancer, it is also scientifically meaningful to incorporate this important biological measure in constructing age prediction intervals. However, commonly used high-dimensional linear regression models for building existing epigenetic clocks implicitly assume that the association between the DNAm profile and chronological age remains the same across different subpopulations, which may not hold due to heterogeneity across different developmental stages, health conditions, and genders[5]. Moreover, constructing prediction intervals from conditional mean regression results in intervals of uniform width, failing to capture individual variability in aging rates or age acceleration, which are influenced by many factors including sociodemographic factors, diet, physical activity, genetics, environmental chemical exposures, and more. Apparently, prediction intervals of the same width, even if statistically valid with nominal coverage, are not able to capture cross-individual heterogeneity in the age acceleration rate across different risk levels.

To bridge this gap, we are interested in looking into subgroup differences across various factors that may affect age acceleration and incorporating such variability into age prediction intervals. First, we propose to build a quantile regression framework for constructing epigenetic clocks. This allows researcher to examine relationships between DNAm levels and different quantiles of the chronological age. Apparently, different age groups may show a different pattern in their methylation levels across CpG sites, calling for age-specific epigenetic clocks. For example, we expect that the same CpG site to have varying effects across different age cohorts. Characterizing such information can help with assessing the impact of DNAm on different age groups or health outcomes. Furthermore, we plan to construct age prediction intervals within the quantile regression framework. For example, to obtain prediction intervals with 90% nominal coverage, we simply fit the conditional quantile function at the 5% and 95% levels to form the corresponding intervals. This method is robust to data with high heteroscedasticity and adaptive to local variability. Besides providing valid coverage in finite samples, the intervals are as short as possible with their length adaptive to individual

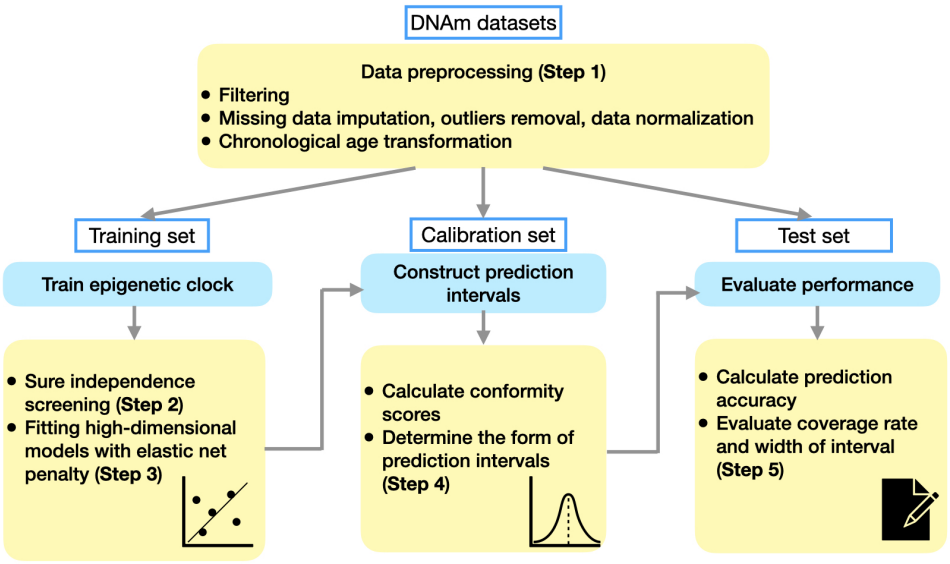


FIGURE 1 Flowchart of the proposed generic pipeline that integrates two key components: (i) training epigenetic clock, (ii) constructing prediction intervals, and (iii) evaluating performance as highlighted in the blue boxes. Details of each step (shown in yellow boxes) will be presented in Section 2.

variability. This enables the identification of individual characteristics that are associated with different levels of uncertainty in their outcome predictions, facilitating tailored interventions.

Throughout this article, we will use children's DNAm data from the Gene Expression Omnibus (GEO) database to illustrate our proposed pipeline for constructing prediction intervals for biological age. Similar to the first-generation epigenetic clocks, we will use chronological age as the outcome of interest. Previous works have adopted a high-dimensional linear regression model with elastic net penalty [14], see for example, the Horvath clock [1], Hannum clock [2], and Levine clock [3]. However, due to the underlying heterogeneity across different age groups, the associations between DNAm levels and age may vary. As a comparison to the conventionally used high-dimensional linear regression model, we will first fit a high-dimensional quantile regression model with the elastic net penalty. Then we will construct prediction intervals and demonstrate their validity as well as statistical efficiency. Our results show that the median regression model outperforms the mean regression model, exhibiting a higher correlation between chronological and predicted ages and a lower median error. Furthermore, our newly proposed approach for constructing prediction intervals not only accommodates subpopulation heterogeneity but is also statistically more efficient than the conventional mean regression-based method.

We describe the generic pipeline of prediction intervals construction for epigenetic clocks in Section 2. Application of the pipeline to establish epigenetic clocks with the children's dataset is given in Section 3. Finally, in Section 4, we present a discussion, outline the limitations, and suggest areas for future research.

2 | METHODS

The overarching goal of this work is to construct prediction intervals for epigenetic clocks with adaptive widths that account for cross-subjects variability. Our proposed generic pipeline involves the following two major components: (i) training high-dimensional predictive models for predicting epigenetic ages, and (ii) constructing valid prediction intervals with the conformal inference methods. A diagram of the proposed generic pipeline is shown in Figure 1.

2.1 | Quantile regression based epigenetic clock

DNA methylation data is typically of ultra-high dimensions since hundreds of thousands of CpG sites are profiled. The whole procedure to construct quantile regression based epigenetic clock can be separated into three steps. In our regression problem, we observe independent and identically distributed samples $(X_i, Y_i) \in \mathbb{R}^q \times \mathbb{R} \sim P$, where $X_i \in \mathbb{R}^q$ is the vector of methylation levels of profiled CpG sites, and $Y_i \in \mathbb{R}$ is the chronological age of subject i for $i = 1, \ldots, n$.

Step 1 (Data preprocessing): As mentioned by [1] and [5], we adopt similar data preprocessing methods. In particular, we firstly integrate data by concentrating on common CpG sites, with several public Illumina DNA datasets relevant to target population. Then we discard any CpG site with more than 10 missing DNA methylation values. For sites with fewer than 10 missing values, we impute with the k-nearest-neighbors approach with the `impute` package in `R`. We study overlapping CpGs that are present on all datasets and carry out the normalization step to ensure that these data are comparable by adapting the `BMIQ R` function from [15] so that it would rescale all probes of each array to match their distribution with the determined gold standard. We also perform a principal component analysis to identify and remove outliers by converting each sample into a Z-score statistic, transforming it to the false-discovery rate, and removing samples falling below a false-discovery rate of 0.2. Optionally, we can transform the chronological age in advance to improve the accuracy of the prediction model. Lastly, the entire dataset is randomly split into training set, calibration set and test set with a proper proportion [16]. Let $\mathcal{D}_{\texttt{train}}$, $\mathcal{D}_{\texttt{cal}}$ and $\mathcal{D}_{\texttt{test}}$ denote the set of sample indices in the training set, calibration set and test set respectively.

Step 2 (Feature screening): Similar to the training procedures for most existing epigenetic clocks, we first adopt the sure independence screening (SIS) method to reduce the number of CpG sites. Since we plan to fit high-dimensional quantile regression models in our subsequent analysis, we choose to work with a model-free generic SIS procedure with fewer and less restrictive assumptions [17]. For example, correlation-based SIS can be summarized as several sketched stages. Firstly, for each feature X_j, compute its correlation with the response variable Y, i.e.,

$$\text{Corr}(X_j, Y) = \frac{\sum_{i=1}^n (X_{ij} - \bar{X}_j)(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_{ij} - \bar{X}_j)^2 \sum_{i=1}^n (Y_i - \bar{Y})^2}},$$

where $\bar{X}_j$ and $\bar{Y}$ are the mean values of the feature X_j and the response variable Y, respectively. Then rank the features based on the absolute values of their correlations, i.e., $|\text{Corr}(X_j, Y)|$. Finally, select the top p features with the highest absolute correlations, where p is a predetermined number of features to retain. Let $\mathcal{S} \subset \{1, 2, \ldots, q\}$ be the set of the index of selected features, and the size of $\mathcal{S}$ is obviously p. To determine the most appropriate feature screening procedure for the analysis of our dataset, we try several popular screening methods that have been implemented in the `MFSIS` package in `R`, such as SIRS [18], DC-SIS [19], Kfilter [20], CSIS [17], Bcor-SIS [21], and WLS [22] to the training set. For each method, we use a 10-fold cross-validation approach to fit regression models predicting chronological age from the selected CpG sites. Within each fold, we optimize the models by selecting the best α and λ parameters and compute the correlation coefficient between predicted values and true values. We average the correlations across all folds for each regression model and choose the feature screening method with the highest average correlation across all models as the best-performing method.

Step 3 (High-dimensional quantile regression): After applying the feature screening method, we then fit a high-dimensional quantile regression model with elastic-net penalty with selected CpGs. It is implemented with the `conquer` package that adopts a convolution-type smoothed quantile regression [23]. Let $\widetilde{X} = (\widetilde{X}_1, \ldots, \widetilde{X}_n) \in \mathbb{R}^{p \times n}$ be the matrix that contains only selected features in Step 2. The estimated coefficient vector $\widehat{\beta}_\tau$ is given by

$$\widehat{\beta}_\tau = \underset{\beta \in \mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^n \rho_\tau \left(Y_i - \widetilde{X}_i^\top \beta \right) + \alpha \left(\lambda \|\beta\|_1 + (1 - \lambda) \|\beta\|_2^2 \right) \right\},$$

where $\rho_\tau(u) = u(\tau - \mathbf{1}_{\{u<0\}})$ is the quantile loss function, $\|\beta\|_1$ and $\|\beta\|_2$ refer to the ℓ_1-norm and ℓ_2-norm separately of the coefficient vector, $\alpha > 0$ is the overall regularization strength, and $0 \leq \lambda \leq 1$ is the regularization parameter that controls the balance between ℓ_1 (Lasso) and ℓ_2 (Ridge) regularization. As a result, the fitted τth conditional quantile of epigenetic age given the methylation levels is formed by $\widehat{f}_\tau(\widetilde{X}) = \widetilde{X}^\top \widehat{\beta}_\tau$, where $\widetilde{X}$ is the vector of methylation level in a new subject.

As a baseline comparison, we also fit a high-dimensional conditional mean regression model with elastic net penalty [14] using the `R` package `glmnet`. The corresponding estimated coefficient vector is the solution to the following optimization problem:

$$\widehat{\beta}_{\texttt{mean}} = \underset{\beta\in\mathbb{R}^p}{\arg\min} \left\{ \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \widetilde{X}_i^\top\beta\right)^2 + \alpha\left(\lambda\|\beta\|_1 + (1-\lambda)\|\beta\|_2^2\right)\right\}.$$

The resulting mean regression based epigenetic clock is given by $\widehat{\mu}(\widetilde{X}) = \widetilde{X}^\top\widehat{\beta}_{\texttt{mean}}$.

2.2 | Uncertainty quantification and interval construction in age prediction

The aforementioned pipeline for constructing epigenetic clocks in Section 2.1 provides point predictions only. However, relying solely on point predictions, without considering the uncertainty or variability, makes it difficult to evaluate the trustworthiness of the predicted age or age-related diseases. Consequently, it is desirable to construct prediction intervals as it quantifies the uncertainty in the prediction and offers a clearer picture of how precise or reliable the age prediction is. While there are various types of conformal prediction methods, we choose to work with the split conformal prediction due to its computational efficiency [24]. Our goal in this section is to construct a level $(1-\gamma)$ prediction interval for biological age prediction.

Ideally, we would like to construct prediction intervals that adapt to the heteroscedasticity in the data. This means the resulting intervals can have wider width in subpopulations with high uncertainty and narrow in those of low uncertainty, providing a more accurate depiction of the confidence in age prediction model. Conformalized quantile regression [16] merges quantile regression and conformal prediction to produce prediction intervals that adapt to the underlying distribution of the data while maintaining rigorous coverage guarantees. This motivates us to integrate high-dimensional quantile regression models to the conformal prediction framework for epigenetic age predictions.

The framework of split conformal prediction will be integrated with epigenetic clocks based on mean regression and quantile regression, respectively. As a continuation of Section 2.1, we use $\widehat{f}_{\gamma/2}(X)$ and $\widehat{f}_{1-\gamma/2}(X)$ to denote the $\gamma/2$th and $(1-\gamma/2)$th condition quantile functions, respectively. We further use $\widehat{\mu}(X)$ to denote the mean regression based epigenetic clock. The split conformal prediction separates the fitting and ranking steps using sample splitting. More specifically, to avoid overfitting, the fitting or training steps in Section 2.1 is done with the training set $\mathcal{D}_{\texttt{train}}$ while the calibration is carried out with $\mathcal{D}_{\texttt{cal}}$.

Step 4 (Residuals and conformity scores): For the mean regression model, we apply $\widehat{\mu}(\widetilde{X})$ to the samples in calibration set and calculate the absolute values of residuals $R_i = \left|Y_i - \widehat{\mu}(\widetilde{X}_i)\right|, i \in \mathcal{D}_{\texttt{cal}}$. This deviation, in the field of epigenetic aging, is interpreted as the magnitude of age acceleration or deceleration, depending on whether Y_i is greater than $\widehat{\mu}(\widetilde{X}_i)$. Among the calculated residuals, we find the $(1-\gamma)$th quantile of the empirical distribution of the absolute values of residuals, denoted by $d_{\texttt{mean}} := d(\mathcal{R}_{\texttt{cal}})$ = the kth smallest value in $\mathcal{R}_{\texttt{cal}} = \{R_i : i \in \mathcal{D}_{\texttt{cal}}\}$, where $k = \lceil (n/2+1)(1-\gamma)\rceil$.

Different from the mean regression models, quantile regression offers a natural framework for constructing prediction intervals possibly formed by by $\widehat{f}_{\gamma/2}(\widetilde{X})$ and $\widehat{f}_{1-\gamma/2}(\widetilde{X})$. To avoid overfitting, we apply these two fitted quantile functions to the calibration set and calculate the conformity scores defined as $S_i = \max\{\widehat{f}_{\gamma/2}(\widetilde{X}_i) - Y_i,\ Y_i - \widehat{f}_{1-\gamma/2}(\widetilde{X}_i)\}, i \in \mathcal{D}_{\texttt{cal}}$. Similarly, among the calculated conformity scores, we find the $(1-\gamma)$th quantile of the empirical conformity score distribution, denoted by $d_{\texttt{quantile}} := d(\mathcal{S}_{\texttt{cal}})$ the kth smallest value in $\mathcal{S}_{\texttt{cal}} = \{S_i : i \in \mathcal{D}_{\texttt{cal}}\}$, where $k = \lceil (n/2+1)(1-\gamma)\rceil$.

Step 5 (Constructing prediction intervals): The prediction interval for the mean regression model is constructed with $C_{\texttt{mean}}(\widetilde{X}_i) = [\widehat{\mu}(\widetilde{X}_i) - d_{\texttt{mean}},\ \widehat{\mu}(\widetilde{X}_i) + d_{\texttt{mean}}]$, for $i \in \mathcal{D}_{\texttt{test}}$. It is clear to see that the conformal prediction technique, if being integrated to the conditional mean regression based epigenetic clock, can be less informative because the resulting prediction intervals are of a fixed width $2d_{\texttt{mean}}$ across all subjects. More importantly, intervals of fixed length do not truly reflect the variation in age acceleration across different stages of human lifespan, or more generally, across subpopulations with different health status, genders, races, and environmental exposures [25].

The prediction interval for quantile regression models is constructed with $C_{\texttt{quantile}}(\widetilde{X}_i) = [\widehat{f}_{\gamma/2}(\widetilde{X}_i) - d_{\texttt{quantile}},\ \widehat{f}_{1-\gamma/2}(\widetilde{X}_i) + d_{\texttt{quantile}}]$, for $i \in \mathcal{D}_{\texttt{test}}$. Different from the prediction interval constructed from mean regression based conformal prediction, this method generates interval with width adaptive to individual DNAm profile. In an ideal case where the fitted conditional quantile functions fit perfectly to the calibration set, and S_i's are zero, the prediction interval is simply formed by $[\widehat{f}_{\gamma/2}(\widetilde{X}_i),\ \widehat{f}_{1-\gamma/2}(\widetilde{X}_i)]$ and the prediction intervals will be formed exactly with the two fitted quantile curves. However, if most S_i's are positive and $q_{\texttt{quantile}}$ is also positive, the resulting interval will be calibrated to a wider interval, and vice versa. As a result, this way of prediction interval construction automatically adjust for both undercoverage and overcoverage.

2.3 | Child-specific methylation-based datasets

We acquire several publicly available DNAm datasets from the GEO database, profiled with Illumina 27K and Illumina 450K array platforms (Table 1). We exclude samples if their chronological ages are missing. These datasets consist of $n = 728$ samples from healthy children with age ranging from 1 month to 216 months. Such a large cohort of children covers the whole age period from 0 to 18 years helps to study the heterogeneous aging pattern precisely throughout childhood. In addition, utilizing these well-established datasets also ensures a rigorous assessment of its performance, thereby reinforcing the credibility and reliability of our proposed pipeline. These datasets will be instrumental in demonstrating the pipeline designed for developing prediction intervals for epigenetic age, as discussed in Section 3.

3 | RESULTS

In this section, we illustrate the proposed pipeline for constructing prediction intervals for epigenetic age with the children's datasets. We first introduce the basic characteristics of the DNAm datasets being used. Then we compare the prediction performances between both mean and quantile regression based epigenetic clocks. Lastly, we assess the validity and statistical efficiency of the prediction intervals constructed from these two types of epigenetic clocks, respectively.

3.1 | Establishment of a child-specific epigenetic clock based on quantile regression

Despite the popularity of conditional mean regression, it is sensitive to outliers and fails to capture heterogeneous relationships between epigenetic age and DNAm profile across different subpopulations. Moreover, in the applications of age and aging-related disease predictions, the underlying heterogeneity across subpopulations may not be fully addressed by inferring the conditional mean. As an alternative modeling approach to the conventional linear regression model, quantile regression allows the relationship to vary across quantiles of the chronological age. This helps to uncover the underlying association patterns which may be quite different between children and adults. Basically, quantile regression provides greater flexibility to identify differing relationships at different parts of the distribution of the outcome variable. It has been reported that the variability in epigenetic age differs across human lifespan[26]. In children and adolescence cohort, for example, the variability in epigenetic age is found to be more drastic in mid-childhood than in toddlers[5]. For this reason, as an analog of existing epigenetic clocks, we propose to fit a high-dimensional quantile regression model with elastic net penalty.

3.2 | Characteristics of the DNA methylation datasets (Step 1)

With the children's datasets, we follow the preprocessing approach illustrated in Step 1 in Section 2.1. Specifically, the gold standard is set to the mean beta value of the largest single dataset (GSE27097), and the age transform function is the same as Function F adopted by[5]. After these steps, our pooled dataset consists of $n = 728$ subjects and $q = 22,233$ CpG sites. Their ages range from 1 month to 216 months (0 - 18 years). Detailed information is summarized in Table 1.

The entire dataset is randomly split into three sub-datasets: 50% for training set, 30% for calibration set and 20% for test set[16]. The training set will be used for training a new epigenetic clock while the test set will be used for evaluating the prediction performance of the newly trained clock, see Section 3.3. The hold out calibration set will be used in Section 3.5 for uncertainty quantification and prediction interval construction.

3.3 | Training a quantile regression based epigenetic clock (Steps 2 to 3)

Following Step 2 in Section 2.1 and according to the results summarized in Table 2, we choose the CSIS method for dimension reduction because it has the highest overall correlation coefficient across different high-dimensional models. The number of recruited predictors is chosen as $p = \lceil n/\log(n) \rceil = 111$. With these selected CpGs and following Step 3 in Section 2.1, we then fit a high-dimensional median regression model with elastic net penalty. As a baseline comparison, we also fit a high-dimensional mean regression model with elastic net penalty. The mixing parameters α are set at 0.56 and 0.49 and the penalty

TABLE 1 Basic information of the children's DNA methylation datasets from the GEO database. After preprocessing, the pooled dataset consists of $n = 728$ samples and $q = 22,233$ CpG sites.

ID	Availability	Methylation array	n	Age range (months)	Mean (SD) age	Citation
1	GSE27097	Illumina 27K	316	43.00-214.00	118.70 (43.79)	[27]
2	GSE32148	Illumina 450K	11	42.00-210.00	136.36 (58.55)	[28]
3	GSE57484	Illumina 27K	8	120.24-127.56	123.11 (2.82)	[29]
4	GSE64495	Illumina 450K	15	27.60-129.60	71.60 (31.95)	[30]
5	E-MTAB-4187	Illumina 450K	81	67.04-196.70	143.17 (40.30)	[31]
6	GSE34257	Illumina 27K	63	1.00-12.00	6.52 (2.87)	[32]
7	GSE36054	Illumina 450K	121	12.00-203.00	59.26 (50.24)	[27]
8	GSE23638	Illumina 27K	13	36.00-192.00	111.69 (51.70)	[33]
9	GSE41037	Illumina 27K	9	192.00-216.00	209.33 (8.72)	[34]
10	GSE52588	Illumina 450K	2	168.00-192.00	180.00 (16.97)	[35]
11	GSE73103	Illumina 450K	89	168.00-216.00	185.39 (9.23)	[36]

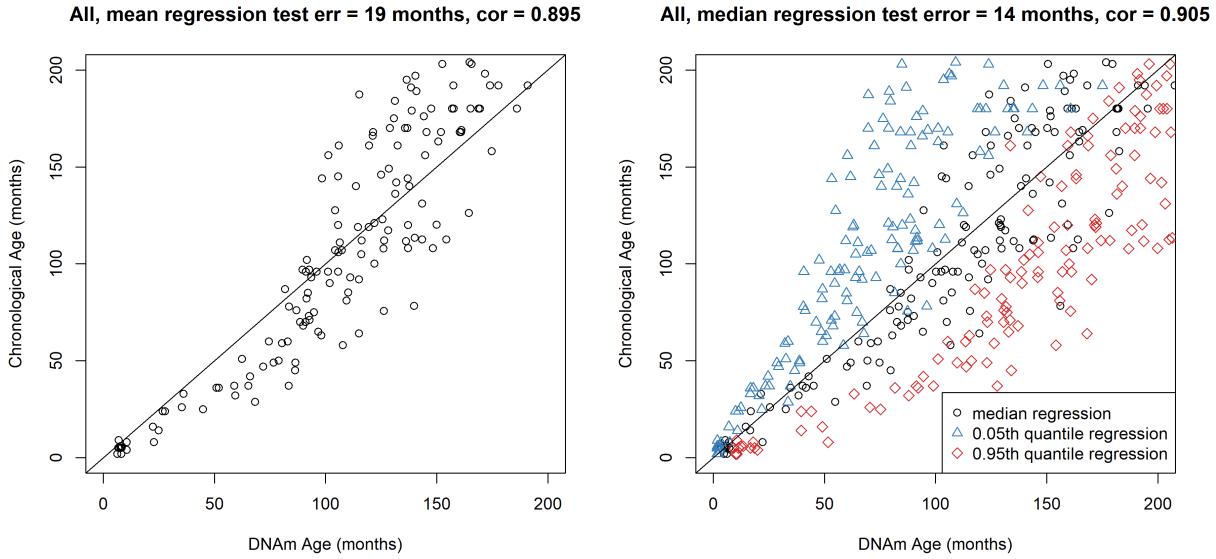
TABLE 2 After applying different screening methods for dimension reduction in the training sets, we fit various regression models and evaluate their performances by calculating the Pearson correlation coefficients. We choose the CSIS method because it has an overall highest correlation coefficient across different regression models.

Screening methods	Mean regression	0.05th quantile regression	0.95th quantile regression	Median regression	Average correlation
SIRS	0.9188	0.8785	0.8885	0.9046	0.8976
DCSIS	0.9264	0.8876	0.8926	0.9081	0.9037
Kfilter	0.8612	0.7841	0.7475	0.7874	0.7951
CSIS	0.9301	0.8849	0.9005	0.9118	**0.9068**
BcorSIS	0.9266	0.8840	0.9007	0.9100	0.9053
WLS	0.8052	0.7128	0.7881	0.7375	0.7609

tuning parameters λ are set at 0.010 and 0.048 in median (`R` function `conquer.cv.reg`) and mean (`R` function `cv.glmnet`) regression models respectively, based on 10-fold cross-validation of the training data.

To evaluate the prediction accuracy, we apply two different training models to predict ages in the test sets and then compare their Pearson correlation coefficients between DNAm age (predicted age) and chronological age, as well as the median absolute difference between DNAm age and chronological age (median error). As shown in Figure 2, the predictive model based on median regression performs slightly better than the mean regression. We find that the median regression model has a correlation coefficient of 0.905 and a median error of 14 months (Figure 2 (b)), while the correlation coefficient and median error of the mean regression model are 0.896 and 19 months, respectively (Figure 2 (a)). This improvement is due largely to the skewness of the age distribution.

In addition to the median regression model, we also fit the 0.05th and 0.95th high-dimensional quantile regression models to examine the lower and upper tails of the chronological age distributions. In our data example, they correspond to toddlerhood and adolescence, respectively. The mixing parameters α are set at 0.36 and 0.48 and the penalty tuning parameters λ are set at 0.021 and 0.010 in 0.05th and 0.95th quantile regression models respectively, based on 10-fold cross-validation of the training data. The correspondences between the estimated DNAm and chronological ages are also plotted in Figure 2 (b). We observe that the blue labels show an upward trend relatively to the diagonal reference line. This is because the DNAm ages represented by these blue triangles are the predicted epigenetic ages with a 0.05th quantile regression model, and they represent the estimated threshold for the 5th percentile of the chronological age, conditional on their DNAm levels. Therefore, the estimated DNAm ages are in general smaller than their true chronological age. Moreover, the deviation from the diagonal line is much smaller for those in lower tail of the chronological age distribution while it gradually becomes larger as we move to the upper tail. A similar deviation pattern can be observed from the red squares, which show the estimated threshold for the 95th percentile of the chronological age. Such a phenomenon reveals different variability in epigenetic age across different developmental stages: it varies much less in toddlerhood than in mid-childhood or adolescence. These results demonstrate the benefit and flexibility of fitting high-dimensional quantile regression models as it provides distributional-level knowledge about the epigenetic age.



(a) Mean regression model

(b) Quantile regression models

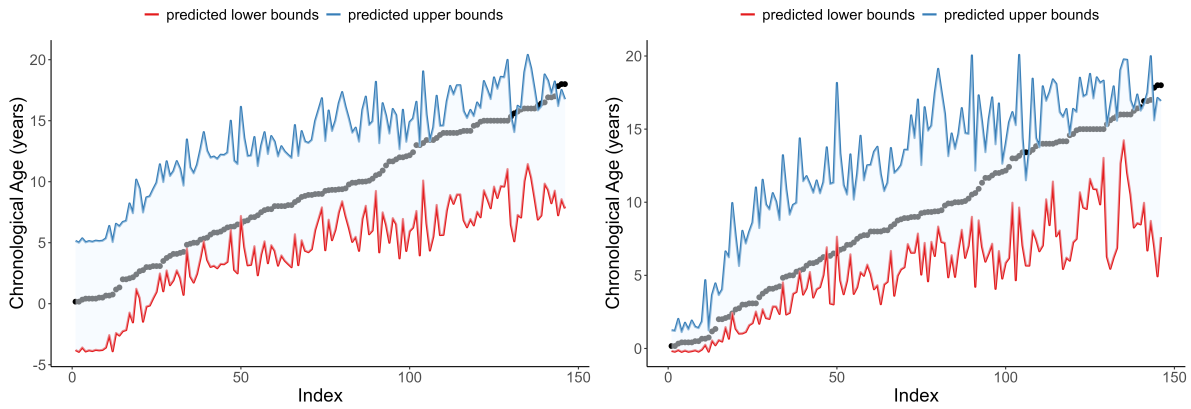
FIGURE 2 Scatterplots for individuals in the test sets (age unit in month). The straight diagonal lines in both plots are the diagonal references for visualizing the consistency between DNAm age and chronological age. The median regression model (black circles) closely aligns along the diagonal line, indicating a strong correlation with chronological age. Meanwhile, the 0.05th quantile regression model (blue triangles) and the 0.95th quantile regression model (red diamonds) display deviations above and below this line, respectively.

3.4 | Constructing prediction intervals and evaluating coverage rate (Steps 4 to 5)

Following Section 2.2, we construct level 90% prediction intervals based on the mean regression and quantile regression models respectively by setting $\gamma = 0.1$. Figure 3(a) shows the prediction intervals derived from the conditional mean regression model. Even though they are statistically valid with coverage rate of 93.8% that is close to nominal level, the width of the intervals remains the same for all subjects. However, it has been found that there is a considerable heterogeneity in physiologic functions and rate of biological aging among individuals of the same chronological age[37]. For example, it has been reported in[5], age acceleration is the greatest in mid-childhood (5-11 years), and we expect to see wider prediction intervals than those in toddlerhood (0-4 years). Consequently, this way of constructing prediction interval overlooks various heterogeneity across different subpopulations or different developmental stages.

In contrast, prediction intervals derived from the quantile regression model are presented in Figure 3(b). With the coverage rate of 91.8% that is also close to the nominal level, the width of the intervals adapts to individuals with different chronological age. Specifically, we observe that the prediction intervals for toddler age group under 4 years old are much narrower than those of the mid-childhood or adolescence according to Figure 4. Such an observation coincides with that has been reported in[5]. Clearly, this approach for constructing prediction intervals provides a more accurate reflection of the heterogeneity in age acceleration across different developmental stages in children.

With both methods attaining the nominal coverage rate, we further compare their statistical efficiency in terms of prediction interval width. Apparently, efficient statistical methods can result in narrower prediction intervals, reflecting greater certainty and confidence in predictions. As shown in Figure 5, prediction intervals derived from the conformalized quantile regression method are narrower than those fixed width intervals produced by conformalized mean regression method: the top of the box on the right, corresponding to the third quartile, is slightly lower than the horizontal line on the left. Consequently, our newly proposed way of constructing epigenetic clocks and their corresponding prediction intervals is not only adaptable to heterogeneity among subpopulations but also statistically more efficient than conventional mean based methods.



(a) Conformalized mean regression (coverage rate=93.8%)

(b) Conformalized quantile regression (coverage rate=91.8%)

FIGURE 3 Prediction intervals for individuals in the test sets (age unit in year). The x-axis denotes subjects indexes and the y-axis represents the true chronological ages. Subject indexes are ordered by chronological ages, as shown in the black dots with an increasing pattern. The blue and red lines denote the upper and lower bounds of prediction intervals, respectively. It clearly shows that the width of intervals remains constant in conformalized mean regression (Panel (a)) while it varies in conformalized quantile regression (Panel (b)).

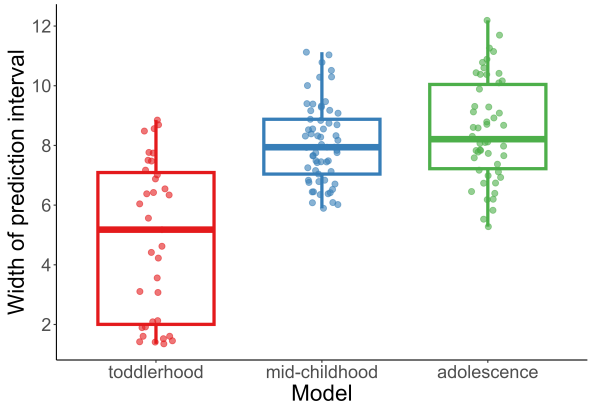


FIGURE 4 Boxplots of width (unit in year) distribution stratified by age groups. The red, blue and green boxplots show the width distributions in toddlerhood, mid-childhood, and adolescence, respectively. The toddler age group has a much narrower prediction intervals than the other two age groups.

3.5 | Prediction intervals with adaptive width better reveals aging pattern

To further investigate the association of the width of prediction intervals with age acceleration, we integrate the average age acceleration (stratified by chronological age) and width of prediction intervals into one plot, as shown in Figure 6. First, we observe that the average acceleration in conformalized mean regression (Panel (a)) is greater than that in conformalized quantile regression (Panel (b)). This implies that the median regression based epigenetic clock fits better into the data, and the corresponding age acceleration better captures the true biological differences rather than measurement error. In addition, both mean and median based methods reveal similar aging patterns: age acceleration is the greatest in mid-childhood, and there is a phenomenon of age deceleration in adolescence. However, in contrast to the fixed width intervals constructed via the conformalized mean regression, the width of prediction intervals produced by the quantile regression has a varying pattern that mimics the variation in age acceleration. Such prediction intervals with widths synchronized to age acceleration provide

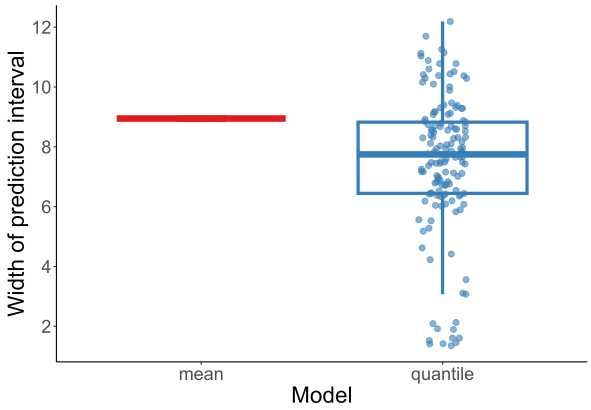
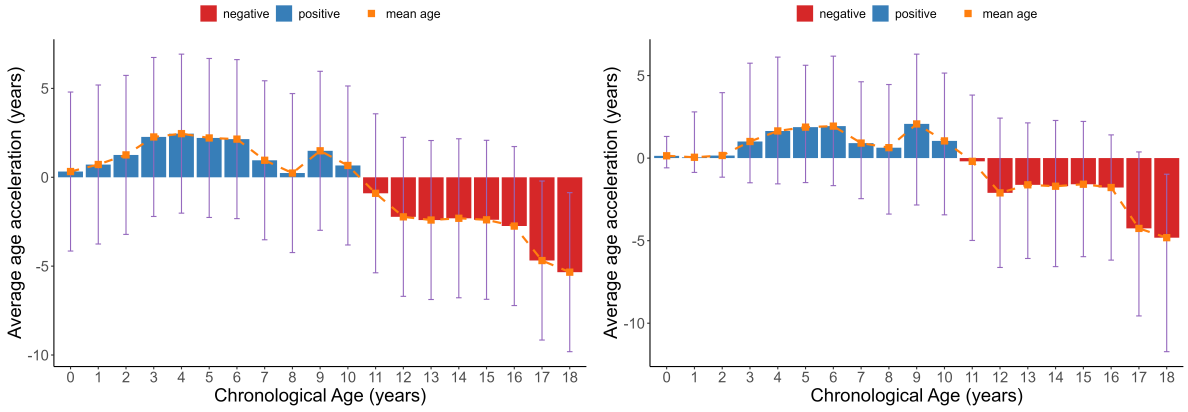


FIGURE 5 Boxplots of width (unit in year) distribution of prediction intervals derived from conformalized mean regression and conformalized quantile regression methods. The red line refers to the fixed width of prediction intervals derived from conformalized mean regression. The blue boxplot denotes the width distribution of prediction intervals derived from conformalized quantile regression. Clearly, conformalized quantile regression method presents a narrower prediction intervals than conformalized mean regression, demonstrating its improved statistical efficiency.

important scientific insights: the possible range of epigenetic age varies greatly for children under different developmental stages. Specifically, for children in their mid-childhood and adolescence, the range of their epigenetic ages may be larger than those in their toddlerhood. In particular, the center of the prediction interval in mid-childhood may be slightly higher than the corresponding chronological age while it is the opposite for those in adolescence.



(a) Conformalized mean regression

(b) Conformalized quantile regression

FIGURE 6 Histogram of the mean value distribution of the age acceleration for individuals in the test sets (age unit in year). The x-axis shows chronological ages, while the y-axis represents the average value in age acceleration. The red column denotes negative value that corresponds to age deceleration, while the blue bar denotes positive average value that corresponds to age acceleration, the orange square indicates the average value, and the purple error bar represents its prediction interval.

4 | DISCUSSION

Although various epigenetic clocks have been developed to estimate the biological age of an individual based on cellular DNA methylation, their models focus on point prediction only[1,2]. However, a single point prediction of biological age is not sufficient for clinical decision-making because it does not convey a level of confidence for predicting the age. In this paper, we extend the conformal inference framework to quantify the uncertainty in age prediction, and integrate it with our newly proposed epigenetic clocks developed with high-dimensional quantile regression model with elastic net penalty. We employ data profiled on Illumina 27K and Illumina 450K array platforms to construct a child-specific epigenetic clock that covers the entire period of childhood (0-18 years old). Our prediction and inference framework has the following advantages over existing age prediction models: (a) it comprehensively reflects the heterogeneity in aging patterns across different subpopulations; (b) it provides prediction intervals that will cover the ground truth with certain probability, and thus can enhance the interpretability in practice; (c) the prediction intervals are adaptive to individual DNAm profile and their varying widths uncover the underlying heterogeneity in age acceleration; and (d) the widths derived from the quantile regression models are generally narrower than those from the mean regression models, and thus are statistically more efficient.

Lifespan dynamics in epigenetic biomarkers are rapidly evolving, with significant implications for aging research, personalized medicine, and public health. Birth to late adolescence, for example, is known to be a tremendously dynamic period of development and growth[38]. It is critical to build biological age prediction models that uncover the underlying population heterogeneity. While mean regression has been widely used due to its simplicity and ease of interpretation, it focuses on predictions about the mean outcome. In contrast, quantile regression provides a more detailed picture by estimating the effects of DNAm across the entire distribution of the chronological age, making it especially useful when understanding the impacts on different developmental stages across lifespan is important. Childhood and adolescence are unique periods of rapid change that is unlikely to mimic adulthood in methylome dynamics[39,40]. Given the differences in the pace of developmental and age-related changes across the life course, developing a unifying prediction model that quantifies the effects of DNAm across different chronological ages will provide a more comprehensive understanding of this dynamic relationship. Our proposed method, when applied to children-specific DNAm datasets, uncovers a clear heterogeneous pattern in different stages of childhood: age acceleration in mid-childhood is much greater than that in toddlerhood, and there is a trend of decceleration in adolescence.

The width of the interval reflects the level of uncertainty associated with the age prediction. Wider intervals indicate higher uncertainty while narrower intervals suggest more confidence in the prediction. This uncertainty can be influenced by various factors such as the inherent noise in the data, the model predictive performance, and the distribution of the data points. According to Figure 2, we observe larger deviations in data points from mid-childhood and adolescence in comparison to those from toddlerhood. This may explain why the resulting intervals in Figure 4 are wider in those two groups to accommodate the increased uncertainty. Such a high variability uncovers the dynamics of DNA methylation levels during mid-childhood, as this period involves significant biological and developmental changes that can be influenced by environmental exposures, genetic factors and lifestyle[3,41]. While a prediction interval with a width of 10 years seems to be less informative in terms of age prediction, it offers insights on a possible range of DNAm ages in a specific developmental stage. In general, epigenetic age acceleration has been used as a measure of biological aging rate and has been linked to various clinical traits such as physical capability and cognitive functioning[42]. An age acceleration of 8 years, for example, should be interpreted differently for a person in mid-childhood versus in toddlerhood. Because for the former, such a large deviation is still within its 90% confidence range. From this perspective, our newly proposed pipeline for constructing prediction intervals provides a normal range in which the observed difference between DNAm age and chronological age may arise from dynamics in that specific developmental stage, rather than a systematic indicator of healthy status.

Although our proposed quantile regression based inference framework can be applied to other types of continuous outcomes such as risk of disease or time to death, we currently do not have data on these outcomes to explore whether there will be new insights when the outcome of interest is not chronological age. For example, using quantile regression, we can estimate how DNAm affect different quantiles of the cardiovascular disease risk distribution. The prediction intervals may have varying lengths for those in lower, median and higher risk groups. In conclusion, constructing prediction intervals based on quantile regression provides a nuanced understanding of how DNAm levels and other demographic features influence outcomes of interest across the entire distribution. By leveraging our newly proposed framework, researchers can gain deeper insights into the dynamics of disease risk and develop interventions that are tailored to the specific needs of different risk groups.

References

1. Horvath S. DNA methylation age of human tissues and cell types. *Genome biology.* 2013;14(10):1–20.
2. Hannum G, Guinney J, Zhao L, et al. Genome-wide methylation profiles reveal quantitative views of human aging rates. *Molecular cell.* 2013;49(2):359–367.
3. Levine ME, Lu AT, Quach A, et al. An epigenetic biomarker of aging for lifespan and healthspan. *Aging (albany NY).* 2018;10(4):573.
4. Lu AT, Quach A, Wilson JG, et al. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging (albany NY).* 2019;11(2):303.
5. Wu X, Chen W, Lin F, et al. DNA methylation profile is a quantitative measure of biological aging in children. *Aging (Albany NY).* 2019;11(22):10031.
6. Belloni A, Chen D, Chernozhukov V, Hansen C. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica.* 2012;80(6):2369–2429.
7. Berk R, Brown L, Buja A, Zhang K, Zhao L. Valid post-selection inference. *The Annals of Statistics.* 2013:802–837.
8. Zhang CH, Zhang SS. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2014;76(1):217–242.
9. Tibshirani RJ, Taylor J, Lockhart R, Tibshirani R. Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association.* 2016;111(514):600–620.
10. Vovk V, Gammerman A, Shafer G. *Algorithmic learning in a random world*. 29. Springer, 2005.
11. Marioni RE, Shah S, McRae AF, et al. The epigenetic clock is correlated with physical and cognitive fitness in the Lothian Birth Cohort 1936. *International journal of epidemiology.* 2015;44(4):1388–1396.
12. Levine ME, Lu AT, Bennett DA, Horvath S. Epigenetic age of the pre-frontal cortex is associated with neuritic plaques, amyloid load, and Alzheimer's disease related cognitive functioning. *Aging (Albany NY).* 2015;7(12):1198.
13. Sibbett RA, Altschul DM, Marioni RE, Deary IJ, Starr JM, Russ TC. DNA methylation-based measures of accelerated biological ageing and the risk of dementia in the oldest-old: a study of the Lothian Birth Cohort 1921. *BMC psychiatry.* 2020;20:1–15.
14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software.* 2010;33(1):1.
15. Teschendorff AE, Marabita F, Lechner M, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics.* 2013;29(2):189–196.
16. Romano Y, Patterson E, Candes E. Conformalized quantile regression. *Advances in neural information processing systems.* 2019;32.
17. Barut E, Fan J, Verhasselt A. Conditional sure independence screening. *Journal of the American Statistical Association.* 2016;111(515):1266–1277.
18. Zhu LP, Li L, Li R, Zhu LX. Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association.* 2011;106(496):1464–1475.
19. Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. *Journal of the American Statistical Association.* 2012;107(499):1129–1139.
20. Mai Q, Zou H. The fused Kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics.* 2015;43(4):1471–1497.
21. Pan W, Wang X, Xiao W, Zhu H. A generic sure independence screening procedure. *Journal of the American Statistical Association.* 2019;114(526):928-937.
22. Zhong W, Liu Y, Zeng P. A model-free variable screening method based on leverage score. *Journal of the American Statistical Association.* 2023;118(541):135–146.
23. Tan KM, Wang L, Zhou WX. High-dimensional quantile regression: Convolution smoothing and concave regularization. *Journal of the Royal Statistical Society Series B: Statistical Methodology.* 2022;84(1):205–233.
24. Lei J, G'Sell M, Rinaldo A, Tibshirani RJ, Wasserman L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association.* 2018;113(523):1094–1111.
25. Zavala DV, Dzikowski N, Gopalan S, et al. Epigenetic Age Acceleration and Chronological Age: Associations With Cognitive Performance in Daily Life. *The Journals of Gerontology: Series A.* 2024;79(1):glad242.
26. Kanherkar RR, Bhatia-Dey N, Csoka AB. Epigenetics across the human lifespan. *Frontiers in cell and developmental biology.* 2014;2:49.
27. Alisch RS, Barwick BG, Chopra P, et al. Age-associated DNA methylation in pediatric populations. *Genome research.* 2012;22(4):623–632.

28. Harris AR, Nagy-Szakal D, Pedersen N, et al. Genome-wide peripheral blood leukocyte DNA methylation microarrays identified a single association with inflammatory bowel diseases. *Inflammatory bowel diseases.* 2012;18(12):2334–2341.

29. Voisin S, Almén MS, Moschonis G, Chrousos GP, Manios Y, Schiöth HB. Dietary fat quality impacts genome-wide DNA methylation patterns in a cross-sectional study of Greek preadolescents. *European Journal of Human Genetics.* 2015;23(5):654–662.

30. Walker RF, Liu JS, Peters BA, et al. Epigenetic age analysis of children who seem to evade aging. *Aging (Albany NY).* 2015;7(5):334.

31. Almstrup K, Lindhardt Johansen M, Busch AS, et al. Pubertal development in healthy children is mirrored by DNA methylation patterns in peripheral blood. *Scientific reports.* 2016;6(1):28657.

32. Khulan B, Cooper WN, Skinner BM, et al. Periconceptional maternal micronutrient supplementation is associated with widespread gender related changes in the epigenome: a study of a unique resource in the Gambia. *Human molecular genetics.* 2012;21(9):2086–2101.

33. Chen Ya, Choufani S, Ferreira JC, Grafodatskaya D, Butcher DT, Weksberg R. Sequence overlap between autosomal and sex-linked probes on the Illumina HumanMethylation27 microarray. *Genomics.* 2011;97(4):214–222.

34. Horvath S, Zhang Y, Langfelder P, et al. Aging effects on DNA methylation modules in human brain and blood tissue. *Genome biology.* 2012;13(10):1–18.

35. Bacalini MG, Gentilini D, Boattini A, et al. Identification of a DNA methylation signature in blood cells from persons with Down Syndrome. *Aging (Albany NY).* 2015;7(2):82.

36. Voisin S, Almén MS, Zheleznyakova GY, et al. Many obesity-associated SNPs strongly associate with DNA methylation changes at proximal promoters and enhancers. *Genome medicine.* 2015;7:1–16.

37. Wagner KH, Cameron-Smith D, Wessner B, Franzke B. Biomarkers of aging: from function to molecular biology. *Nutrients.* 2016;8(6):338.

38. McEwen LM, O'Donnell KJ, McGill MG, et al. The PedBE clock accurately estimates DNA methylation age in pediatric buccal cells. *Proceedings of the National Academy of Sciences.* 2020;117(38):23329–23335.

39. Barker ED, Walton E, Cecil CA. Annual Research Review: DNA methylation as a mediator in the association between risk exposure and child and adolescent psychopathology. *Journal of Child Psychology and Psychiatry.* 2018;59(4):303–322.

40. Cecil CA, Neumann A, Walton E. Epigenetics applied to child and adolescent mental health: Progress, challenges and opportunities. *JCPP advances.* 2023;3(1):e12133.

41. Teschendorff AE, Relton CL. Statistical and integrative system-level analysis of DNA methylation data. *Nature Reviews Genetics.* 2018;19(3):129–147.

42. Faul JD, Kim JK, Levine ME, Thyagarajan B, Weir DR, Crimmins EM. Epigenetic-based age acceleration in a representative sample of older Americans: Associations with aging-related morbidity and mortality. *Proceedings of the National Academy of Sciences.* 2023;120(9):e2215840120.