Title
Site Selection Methods and Applications of the Epigenetic Pacemaker (EPM) Clock

Author
Huang, Huiling

Publication Date
2020

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Site Selection Methods and Applications of the Epigenetic Pacemaker (EPM) Clock

A thesis submitted in partial satisfaction
of the requirements for the degree
Master of Science in Bioinformatics

by

Huiling Huang

2020

ABSTRACT OF THE THESIS

Site Selection Methods and Applications
of the Epigenetic Pacemaker (EPM) Clock

by

Huiling Huang
Master of Science in Bioinformatics
University of California, Los Angeles, 2020
Professor Matteo Pellegrini, Chair

Various epigenetic clocks have been constructed using DNA methylation data, using regression models to estimate age from DNA methylation patterns. To overcome the constraints imposed by these epigenetic clocks, the Epigenetic Pacemaker (EPM) clock is built to predict an individual's epigenetic state in an unbiased non-linear manner. The EPM clock models the initial methylation value and rate of change in methylation at each methylation locus, enabling an intuitive interpretation of coefficients of selected sites. Since the EPM model is computationally heavy, selecting informative loci in the preprocessing step is necessary. We selected model sites using either a novel randomized ridge regression selection method or the Pearson Correlation Coefficient (PCC) method. The PCC metric achieved higher performance and was used as the site selection method when applying EPM clock to a schizophrenia data set. In this data set, age acceleration predicted by EPM model was positively correlated with schizophrenia status and sex as a male. By experimenting with different EPM models, we conclude that a full model using all samples to build an EPM model generates stable epigenetic state predictions. Building EPM model using more sites

with higher PCC values correlated with phenotype traits is more informative.

The thesis of Huiling Huang is approved.

Eric J. Deeds

Roy Wollman

Matteo Pellegrini, Committee Chair

University of California, Los Angeles

2020

To my parents & grandparents,

my dear friends,

and ZYL, BJYX.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

Epigenetics refers to various types of covalent modification to chromatin that affect gene expression without modifying an individual's DNA sequence. DNA methylation is a particular type of epigenetic control, where a cytosine in a CpG dinucleotides is methylated [1]. The status of the epigenome changes overtime responding to environmental factors [2] yet maintaining a relatively stable representation of an individual's current physiological condition. As a result, it is not surprising that many studies have shown that DNA methylation changes as an organism ages. This gives rise to the notion of DNA methylation age, also known as epigenetic age, which is predicted by DNA methylation data and is a biomarker of aging. As a biomarker, epigenetic age may better predict health than chronological age does [3]. Down syndrome, a disease that would raise the risks of many chronic diseases which are typically associated with older age, exhibits accelerated epigenetic aging in blood and brain tissue [4]. This molecular evidence suggests Down syndrome individuals' functional capability may decrease faster than average.

To calculate epigenetic age, various epigenetic clocks have been constructed. Inputting DNA methylation data of an individual into an epigenetic clock generates the individual's predicted epigenetic age. Most epigenetic clocks use a penalized regression approach to select CpG sites associated with chronological age and build a linear model to predict epigenetic age. The model is a linear combination of the methylation values at selected CpG sites and the weights corresponding to each site generated by regression. The Horvath epigenetic clock, the first multi-tissue epigenetic age predictor, uses elastic net regression to build

the model, yielding 353 CpG sites [1]. The Hannum epigenetic clock, another widely used epigenetic clock, also applies elastic net regression to the methylation data, but combined with bootstrap approaches, yielded only 71 CpG sites [5]. While these two clocks both predict epigenetic ages, there are some limitations of the penalized regression approach, including the biological interpretation of weights and sites selected, and the biological explanation of why minimizing the objective function of the model would yield the epigenetic age.

As there are more and more epigenetic clocks built using machine learning algorithms that generate predictions, what these epigenetic clocks teach us about the biology of aging becomes a more vital question [6]. A recently constructed epigenetic clock called the epigenetic pacemaker (EPM) clock answers this biological question using a different approach and addresses part of the limitations of the penalized regression model. EPM parametrizes epigenetic state, initial methylation value at each CpG site, and the rates of change over time at each CpG site [7], enabling a nonlinear prediction of epigenetic state. The rates of change in EPM are comparable to the weights in the penalized regression model, and the rates of change are biologically interpretable.

However, site selection appears to demand more attention when building an EPM clock. The first problem of site selection is in the preprocessing step. Since EPM uses a conditional expectation-maximization algorithm to fit the model [7], which is more computationally heavy than regression models, site selection before fitting the EPM is necessary. The original EPM retains the 1000 best CpG sites combining different selection methods, including variance, covariance, and Pearson correlation [7]. We propose to consider penalized regression and retain more potential CpG sites. The second site selection problem is during the fitting step. We found that after fitting an EPM model using cross-validation, many of the rates of change returned by the EPM model equal to zero. Removing these zero-rate sites might help improve the model.

After proper selection of sites, an EPM clock is built with DNA methylation values of these selected CpG sites. Then the EPM model can be applied to real data sets and

predict epigenetic state. To determine the practicality of the EPM model, how deviation of individual epigenetic age from the general trend (age acceleration) correlates with individual phenotype traits is studied. If an EPM model performs well, individuals with certain traits will have higher age acceleration than others. For example, we would expect that patients of certain diseases have higher age acceleration than controls. To research on how to build a good EPM model, different model-fitting methods and various sets of selected sites were experimented. To compare and evaluate different models, how well age acceleration predicted by a model is explained by phenotype traits is quantified by different methods.

CHAPTER 2

Methods

2.1 Data Acquisition

2.1.1 Data for Studying Site Selection Methods

14 Illumina 450k methylation data sets were collected from the Gene Expression Omnibus (GEO) repository. Some of the data sets contain phenotype data of the individuals. All methylation data were quantile normalized, and samples with excessive missing methylation data or phenotype data were dropped. 1143 samples were retained for studying site selection methods.

2.1.2 Data for Applying EPM clock

From the Gene Expression Omnibus (GEO) repository, an Illumina 450k DNA methylation data set for schizophrenia patients and controls (Series GSE84727) was used for studying applications of the EPM clock. There are 847 whole blood-derived DNA samples, consisting of 414 schizophrenia cases and 433 controls. Samples without available chronological age were dropped. 665 samples were retained for further study, consisting of 260 schizophrenia cases and 405 controls. There were more male cases (480 cases) than females cases (185 cases) among the retained individuals. The chronological age range of these 665 samples is from 18.3 years old to 80.7 years old.

2.2 Epigenetic Pacemaker (EPM) Model

2.2.1 Model Formulation

The EPM model is constructed by the adaption of the universal pacemaker of genome evolution to the epigenetic setting [7]. It is the first epigenetic clock that parametrizes rates of change in methylation at each site and relaxes the time-linear constraint.

Let m be the number of individuals, n be the number of methylation sites for each individual. For each individual $j, j = 1, 2, \ldots, p$, the methylation value at site $i, i = 1, 2, \ldots, q$, is denoted as $m_{i,j}$. Each individual j has a corresponding state s_j, which is the individual's epigenetic state. For each methylation site i, m_i^0 denotes the initial methylation value and r_i denote the rate of change in methylation. The EPM model states that $m_{i,j} = m_i^0 + r_i s_j$[7], where each site's methylation value changes linearly at a constant rate as the epigenetic state changes. The observed methylation value at site i for individual j is denoted as $\hat{m_{i,j}} = m_i^0 + r_i s_j + \varepsilon_{i,j}$ [7], where $\varepsilon_{i,j}$ is a normally distributed error term. The EPM model optimizes m_i^0, r_i, s_j by minimizing sum of squared error term utilizing conditional expectation maximization (CEM) algorithm. CEM iteratively fixes s_j and optimize m_i^0, r_i, then fixes the optimized m_i^0, r_i to optimize s_j, until the improvement of the model is below certain threshold. s_j is initialized as the individual's chronological age, then CEM starts the iterations. When the iterations stop, s_j are the predicted epigenetic states by the EPM model.

2.2.2 Cross Validation

The model can either be fitted once using all the data points, or it can be fitted using cross-validation.

When performing a cross-validated EPM model, for each site, we took the average of m_i^0, r_i generated on training data across different folds. We used this set of initial methylation

values and rates of change in methylation to optimize the epigenetic state as the prediction generated by the cross-validated EPM model.

2.3 Site Selection Methods

2.3.1 Methylation Site Selection in Preprocessing Step

The computational cost of CEM algorithm goes up dramatically as the number of sites used increases. Since CEM is computationally heavy, it is necessary to preprocess the 450,000 sites in the methylation data to a smaller subset that are potentially significant when predicting epigenetic age. When exploring site selection methods, 20 percent of the samples were used for site selection, and the remaining 80 percent were used for fitting the EPM model and getting epigenetic age predictions. Variance, covariance, and Pearson correlation are some metrics to perform site selection [7].

Two site selection methods in the preprocessing step were performed and compared. First, we used a novel approach, randomized ridge regression to select methylation sites that are potentially significant. Iteratively, we randomly sample a subset of the data and perform ridge regression on methylation data and chronological age. For each iteration of ridge regression, we recorded the sites that have a weight above a certain threshold. After some iterations, a number of sites had been recorded and some sites appeared multiple times. Sites that appeared above a certain number of times were selected. The number of iterations, threshold to record sites during each iteration, and the threshold of the count of sites after all iterations are parameters that could be tuned to obtain the desired amount of selected methylation sites. Second, the Pearson Correlation Coefficient (PCC) metric is used for site selection in the preprocessing step. For all 450,000 sites in the data set, the PCC score is calculated between the methylation value at each site for all individuals and all individual's chronological age for studying different site selection methods. Sites with highest PCC scores are selected.

2.3.2 Methylation Site Removal Method in Model Refitting Step

After selecting potential sites in the preprocessing step, we use these sites to fit the cross-validated EPM model. The returned rates of change in methylation, r_i, are examined, and we found that some of the r_i are close to zero. This means that the methylation value at these sites does not change with the epigenetic state, and they are thus unable to capture the features of the epigenetic state. We proposed to remove these sites that have rates of change close to zero to enhance the simplicity of the model and examine whether removing these sites would retain the captured signals, or even improve the performance of the model.

To remove the sites that have rates of change close to zero, we iteratively fit the cross-validated EPM model until some stopping criteria. A fixed threshold is picked. We first fit a cross-validated EPM model using the starting sites selected from the preprocessing step, then we remove the sites that have rates of change under the threshold. Then we refit the EPM model using the selected sites from the previous step until there are no sites that have a rate of change under the threshold. The threshold is a hyperparameter, and different thresholds are tested. Epigenetic states predicted by models with different thresholds are evaluated. We pick the threshold that generates the model with the highest accuracy when the sites removal step ends iterations.

Model accuracy is evaluated based on the coefficient of determination (R-squared) between predicted epigenetic state and the trendline of all predicted epigenetic states.

2.4 EPM Model Applications

To apply the EPM model to real data sets, there are different options in the model fitting step. Besides, different sets of selected sites can be experimented to build the EPM model.

2.4.1 Options in Model Fitting

1. **Full Model**

   An EPM model can be fitted once using all samples, denoted as a Full Model. The epigenetic state prediction is directly within the results of the fitted model. Every sample has an epigenetic state prediction.

2. **Cross-Validated Full Model (CV Full Model)**

   An EPM model can be fitted with cross-validation using all samples, denoted as Cross-Validated Full Model. The epigenetic state prediction is directly within the results of the fitted model. Every sample has an epigenetic state prediction.

3. **Cross-Validated Model Using Train-Test Split (CV TT Model)**

   Samples can be divided into a training and a testing set. An EPM model can be fitted with cross-validation on the training set, then the epigenetic state is predicted using the fitted model on the testing set. Only samples in the training set have epigenetic state prediction.

2.4.2 Options in Site Selection

After choosing a site selection method, the next step is to determine which set of sites should be used in fitting the EPM model. Typically, sites correlated with age are used for further fitting an EPM model, as described in **Section 2.3 Site Selection Methods**. However, different sets of sites can also be experimented to fit the EPM model, not confined to sites correlated with age.

Below are the proposed combinations of sites for the schizophrenia data set, considering available phenotype data in the data set GSE84727.

1. **Sites Correlated with Chronological Age**

M1.1	Full Model with Sites Correlated with Chronological Age
M1.2	Full Model with Sites Correlated with Chronological Age, Sex, and Disease
M1.3	Full Model with Sites Correlated with Chronological Age and Sex
M2.1	CV TT Model with Sites Correlated with Chronological Age
M2.2	CV TT Model with Sites Correlated with Chronological Age, Sex, and Disease
M2.3	CV TT Model with Sites Correlated with Chronological Age and Sex

Table 2.1: Proposed EPM Models

2. **Sites Correlated with Chronological Age, Sex, and Disease**

3. **Sites Correlated with Chronological Age and Sex**

In each combination, sites correlated with a specific phenotype are derived using the site selection method. Then, the union of the sites correlated with all phenotypes in each combination is used to fit an EPM model. Take combination **3.** as an example, sites correlated with age, and sites correlated with sex are found independently using the site selection method. Then, the union of the sites correlated with age and sites correlated with sex is used for further model fitting.

2.4.3 Proposed EPM Models

When experimenting with different model-fitting options, the CV Full Model yields almost the same results as the Full Model for the schizophrenia data set GSE84727. As a result, only the Full Model and the CV TT Model are considered when proposing EPM models. Combining all the options in model fitting and options in site selection, six models were proposed and labeled as the following.

Each proposed EPM model was applied to the schizophrenia data set GSE84727, yielding different epigenetic state predictions. A conventional linear model using the same set of sites

was also applied to the data set. The results from the EPM model and the linear model were compared within the proposed models. The results from all the proposed EPM models on the same data set were then evaluated and compared.

2.4.4 Evaluation of the EPM Models

To utilize EPM models for biological implications, we are interested in how individual epigenetic state prediction deviates from the trend of the overall population, and how such deviation could be explained by different phenotype traits. The following 2 evaluation methods both aim for exploring the relationship between individual epigenetic state deviation and phenotype traits.

2.4.4.1 Correlating Age Acceleration with Phenotype Traits

Age acceleration is defined as the difference between epigenetic state prediction and the overall trend line of epigenetic state predictions. For this data set, a square root function of epigenetic state predictions on chronological age is fitted on all samples. For each sample, expected epigenetic state prediction can be calculated from the fitted square root function using know chronological age. Age acceleration of this sample can then be obtained by subtracting expected epigenetic state prediction from epigenetic state prediction.

Since the phenotype traits in the schizophrenia data set are binary (sex and disease status), a Mann-Whitney U test was performed on age acceleration grouped by different values of a specific phenotype trait. If the p-value of one such test is significant, one can conclude that age acceleration captures the difference of the phenotype trait. The more age acceleration can differentiate among categories of phenotype traits, the better the EPM model is.

2.4.4.2 Modeling Epigenetic State Prediction Using Chronological Age and Phenotype Traits

Another way of studying how deviation of individual epigenetic state prediction from general trend can be explained by phenotype variation is to model epigenetic state prediction using chronological age and phenotype traits, in this data set, sex, and disease status. Intuitively, epigenetic state prediction is heavily correlated with chronological age. What we are interested in is how the residuals after regressing out chronological age are explained by phenotype traits.

To study the relationship of the residuals and phenotype traits, a generalized additive model (GAM) is built for epigenetic state prediction on chronological age, sex, and disease status. Since sex and disease status are binary variables, they were included in the model as conventional linear covariates. The continuous variable, chronological age, was specified with a smooth function, allowing non-linear relationship with epigenetic state prediction. The significance of parametric coefficients could be observed by looking at the summary table of the GAM model in R.

CHAPTER 3

Results

3.1 Site Selection Methods

3.1.1 Comparing Site Selection Methods in Preprocessing Step

In the preprocessing step, either Randomized Ridge Regression or Pearson Correlation Coefficient (PCC) were applied to select sites. For each site selection method, a different number of selected starting sites were tested to fit a cross-validated EPM model. Overall, PCC achieved higher performance than Randomized Ridge Regression. Among starting sites ranging from 450 to 1600, the EPM model using the PCC site selection method achieves higher performance than the model fit by Randomized Ridge Regression. For the PCC method, as the number of starting sites increases, the model performance first increases then decreases. For the Randomized Ridge Regression method, the model performance increases as the number of starting sites increases.

However, the R-squared, which is the metric for model accuracy, overall doesn't vary much with different methods and different starting sites. For the PCC method, the maximum R-squared achieved is 0.868 with 450 starting sites, and the minimum R-squared achieved is 0.848 with 1565 starting sites. The difference between the minimum and maximum for PCC is 0.02. For the Randomized Ridge Regression method, the maximum R-squared achieved is 0.830 with 1565 starting sites, and the minimum R-squared achieved is 0.793 with 465 starting sites. The difference between the minimum and maximum for Randomized Ridge Regression is 0.037. For all the built models, the difference between the minimum and

Figure 3.1: Site Selection Methods in Preprocessing Step

maximum R-squared is 0.075.

In Figure 1, each dot represents the R-squared between predicted epigenetic age and trendline of all predicted epigenetic ages, labeled on the y-axis, of an EPM model fitted with the number of starting sites labeled on the x-axis. For Randomized Ridge Regression method (blue dots), number of starting sites are tested from 465 sites to 1565 sites. For PCC method (orange dots), number of starting sites are tested from 300 to 1565 sites. Randomized Ridge Regression method (blue line) reached a maximum r-squared of 0.830 with 1565 sites. PCC method (orange line) reached a maximum r-squared of 0.868 with 450 sites.

3.1.2 Applying Site Removal Method in Model Refitting Step

After site selection in the preprocessing step, we applied a site removal method in the model refitting step. We plotted the distribution of rates of change in the preprocessing step for each preprocessing method. Applying PCC in the preprocessing step produces a bimodal shape of rates of change with few rates around zero (Figure 2A), while Randomized Ridge

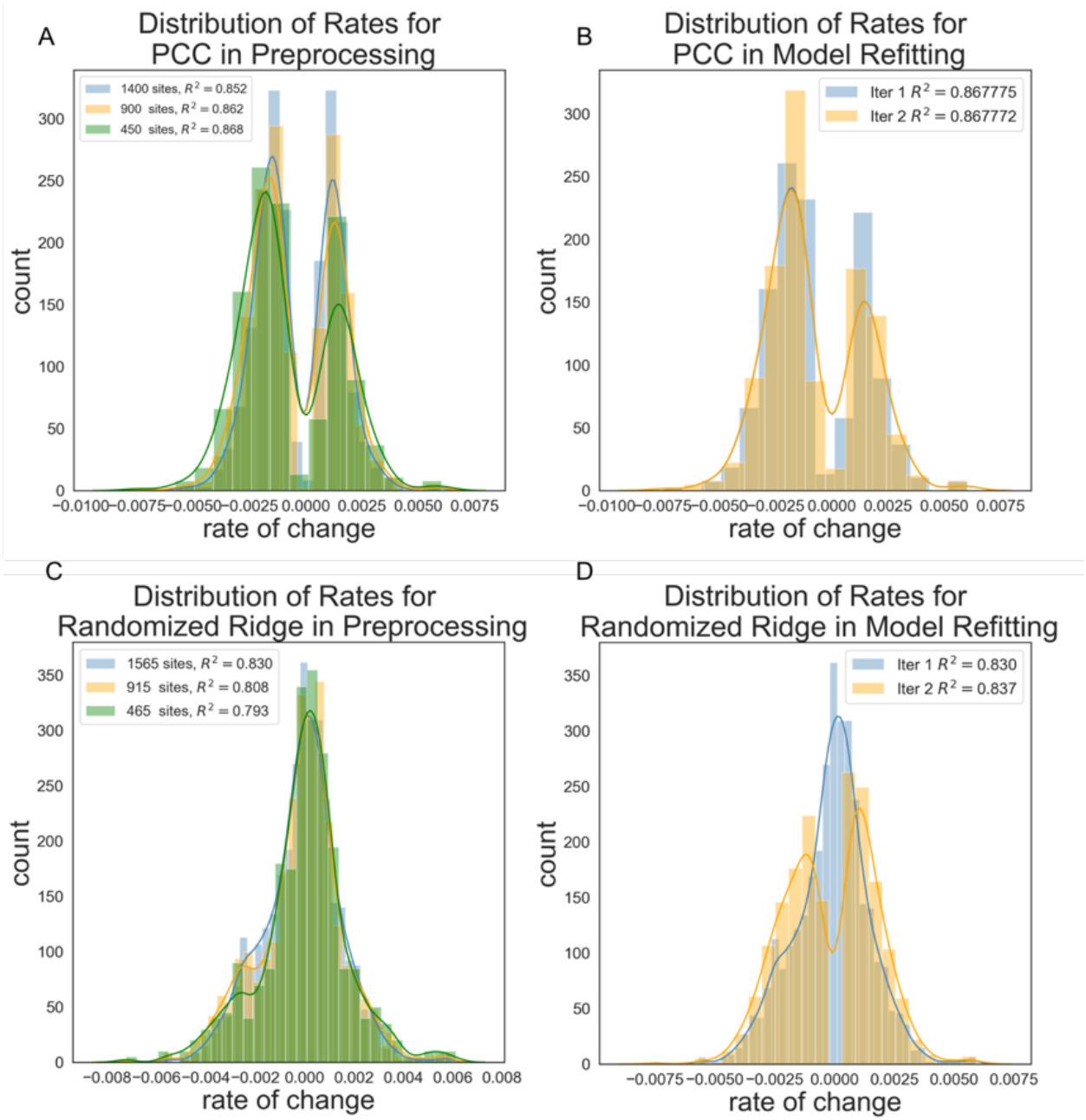


Figure 3.2: Distribution of Rates of Change when Selecting Sites

Regression generates a normal distribution with many rates around zero (Figure 2C). We applied the site removal method in the model refitting step to the best model for each preprocessing method to observe whether site removal methods increase model accuracy.

For the PCC method with preprocessing, the 450-site EPM model has the highest R-squared. After applying site removal on the 450-site PCC model, the shape of the distribution of rates didn't change and the model accuracy didn't change either(Figure 2B). The site removal method during the model fitting step is not necessary for PCC preprocessing.

For the Randomized Ridge Regression method preprocessing, the 1565-site EPM model has the highest R-squared. The site removal method during the model fitting step efficiently removes the sites with rates close to zero, from 1565 sites to 1056 sites. The distribution of rates changed form a normal shape to a desired bimodal shape (Figure 2D). R-squared increased but remained lower than the r-squared of the EPM model with sites preprocessed using the PCC method.

3.1.3 Stability of EPM Model with Site Removal Method

The EPM model remains stable when applying the site removal method on Randomized Ridge Regression preprocessed sites in the model refitting step. Predicted epigenetic states of the EPM model with starting sites from the preprocessing step and predicted epigenetic states after site removal method on these starting sites are closely correlated.

In Figure, predicted epigenetic ages of EPM model with 1565 starting sites preprocessed by Randomized Ridge Regression and predicted epigenetic ages with 1056 sites after site removal method are closely correlated.

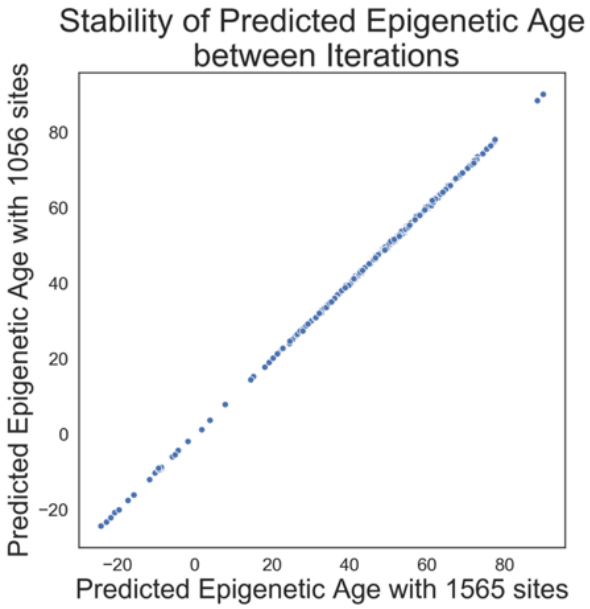


Figure 3.3: Stability of Predicted Epigenetic State between Iterations

3.2 Application of the EPM Model to Real Data Set

3.2.1 Experimenting Different EPM Models

Based on exploration in previous sections, the PCC method was used to select sites based on PCC between methylation values and specific phenotype traits, where only sites with an absolute value of PCC above a certain threshold were retained. The number of selected sites and the threshold is defined by the user.

For each proposed EPM model, a corresponding linear model was also built to compare with the EPM model. The conventional linear model uses the same selected sites as in the proposed EPM model.

Two evaluation methods were used as proposed, correlating age acceleration with phenotype traits, and modeling chronological age with predicted epigenetic age and phenotype traits using the GAM function. Detailed summary table and fitting curves of the GAM functions generated by R are in the supplement figures section.

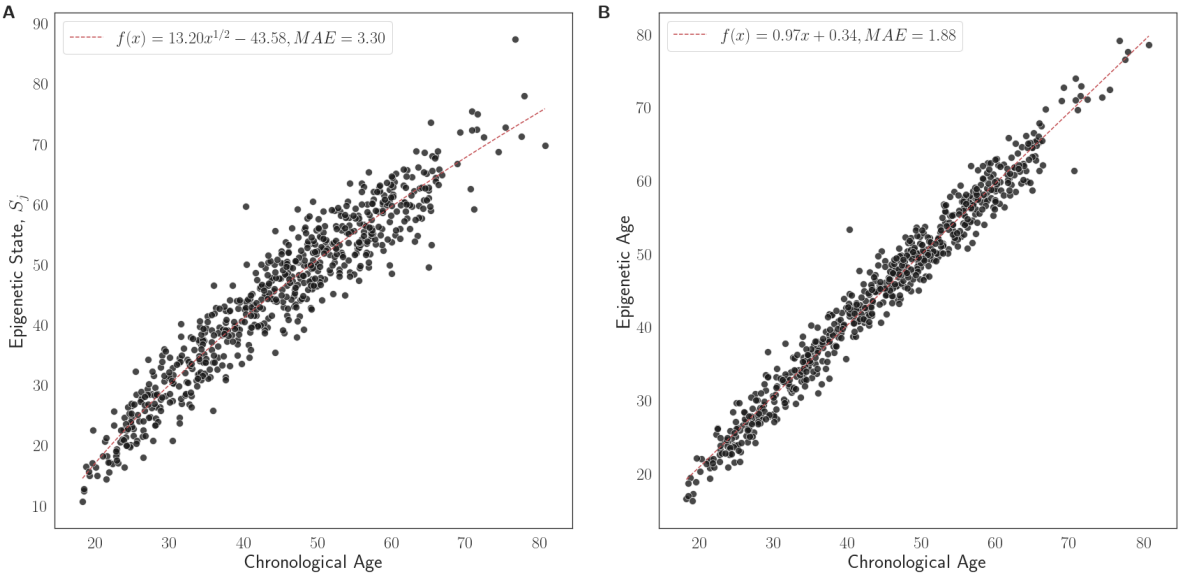


Figure 3.4: Chronological age and predicted epigenetic state of **A) M1.1:** Full Model with Sites Correlated with Chronological Age, and **B)** corresponding linear model

- **M1.1: Full Model with Sites Correlated with Chronological Age**

  For a full model, methylation data of all 665 samples were included when selecting sites. A PCC threshold of 0.58 was chosen to obtain 206 selected sites correlated with chronological age. In 3.4, the corresponding linear model demonstrated a tighter pattern around the predicted trend line of epigenetic state.

  The overall residuals of the linear model was significantly smaller than the EPM model. However, only the EPM model significantly separated phenotype traits with age acceleration, while the linear model did not. 3.5a and 3.5c revealed how well EPM age acceleration differentiated categories of disease status and sex respectively, with corresponding p-value of 0.00003 and 0.02879. In general, samples with schizophrenia are having higher EPM age acceleration, and males are having higher EPM age acceleration than females.

  The results were consistent with the GAM model built on chronological age with epigenetic state, sex, and disease status. In the GAM model using EPM epigenetic state

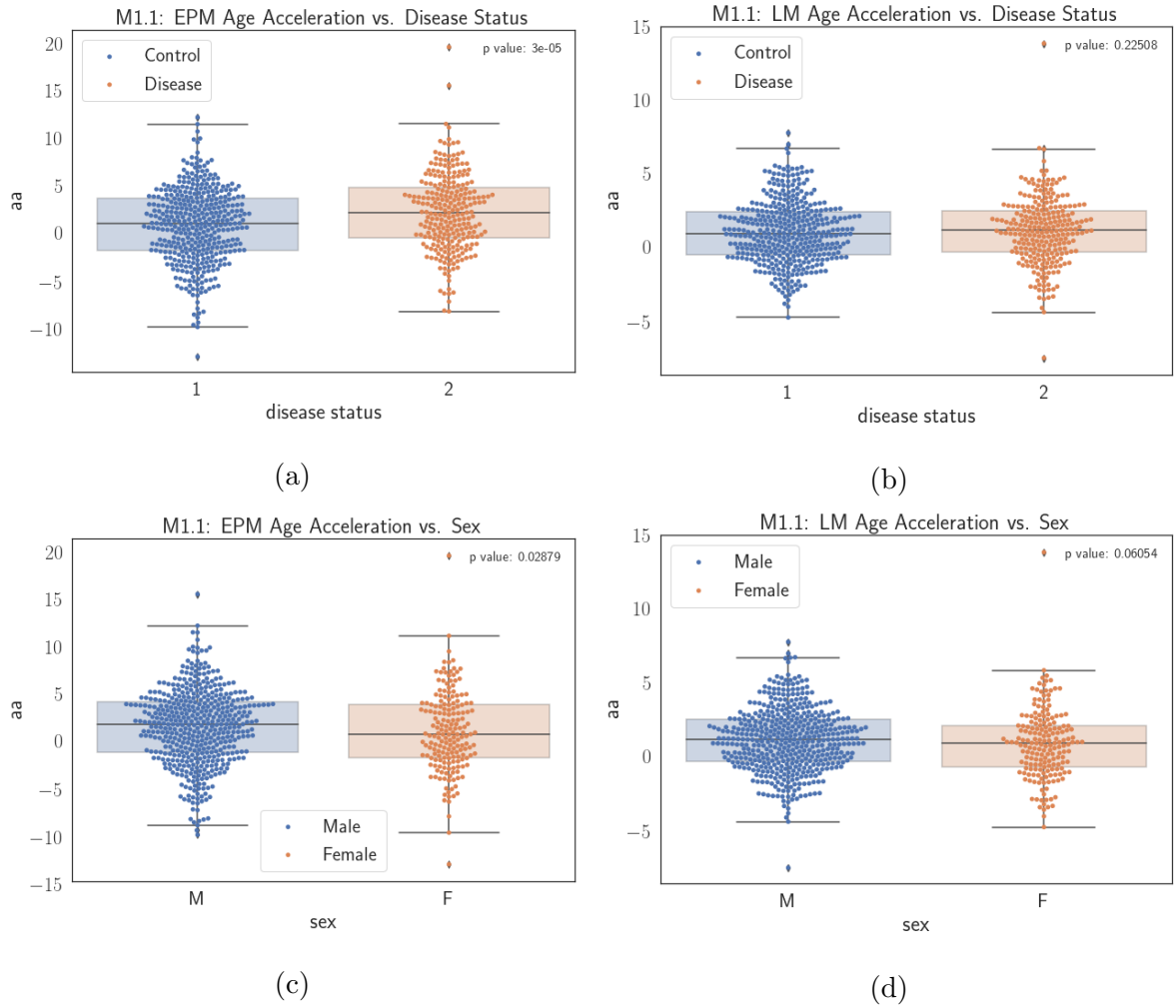


Figure 3.5: Age acceleration and phenotype traits of **M1.1:** Full Model with Sites Correlated with Chronological Age.

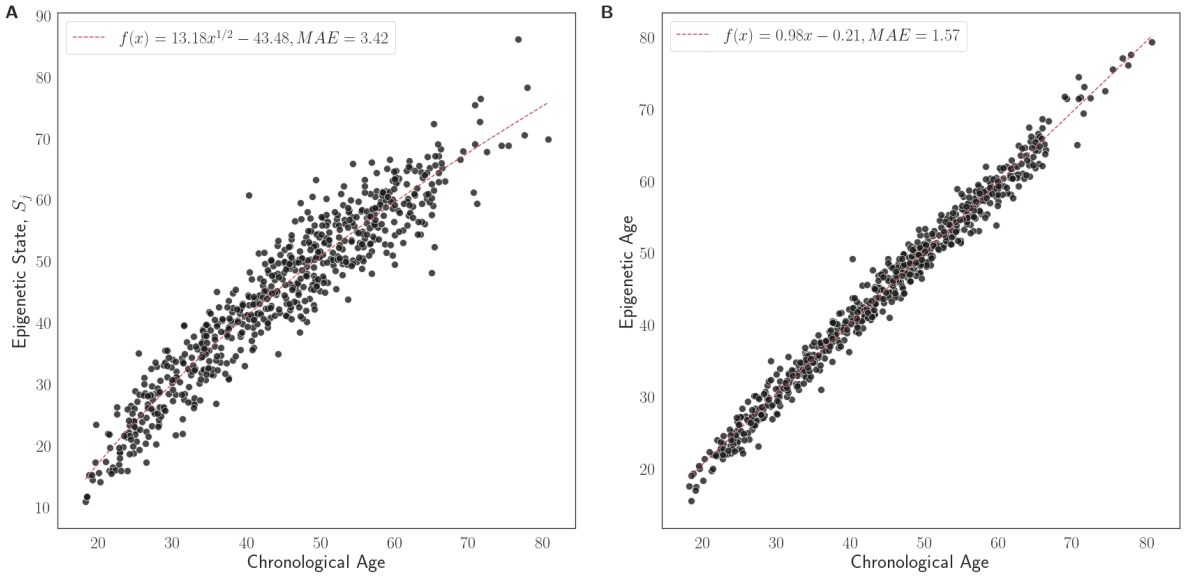


Figure 3.6: Chronological age and predicted epigenetic state of **A) M1.2:** Full Model with Sites Correlated with Chronological Age, Sex, and Disease, and **B)** corresponding linear model

predictions, the parametric coefficients of chronological age, intercept, and disease status were significant. The covariate sex was slightly significant. While in GAM model using linear model epigenetic state predictions, only the parametric coefficients of chronological age and intercept were significant. Both sex as a male and disease status as positive are having positive effects on epigenetic state.

- **M1.2: Full Model with Sites Correlated with Chronological Age, Sex, and Disease**

  M1.2 was also a full model. More sites correlated with different phenotype traits were included in the site selection process before building the EPM model, aiming to boost performance. The 206 selected sites correlated with chronological age obtained in M1.1 was the same across M1.1, M1.2, and M1.3. The same PCC threshold of 0.58 was chosen to obtain sites correlated with sex, and there were 90 sites selected. To keep the same number of sites correlated with disease status and sites correlated with

sex, a PCC threshold of 0.348 was chosen to obtain 91 sites correlated with disease status. There were no overlapping sites, and a total number of 387 sites were used to train the EPM model and the corresponding linear model. In 3.6, the corresponding linear model demonstrated a much tighter pattern around the predicted trend line of epigenetic state.

Although the overall pattern of trend lines of the EPM model and the corresponding linear model in 3.6 were not so different from the ones in 3.4, the correlation between either EPM age acceleration or linear model age acceleration and phenotype traits were significantly lower, as shown in 3.7. None of the combinations between age acceleration and phenotype trait had significant p-value, which means that M1.2 did not perform well at all.

The results were relatively consistent with the GAM model built on chronological age with epigenetic state, sex, and disease status. In the GAM model using linear predictions, only the parametric coefficients of chronological age and intercept were significant. In the GAM model using EPM epigenetic state predictions, the parametric coefficients of chronological age, sex, and intercept were significant. This was understandable since the p-value of the correlation between EPM epigenetic age acceleration and sex is close to 0.05, almost significant.

Combining the p-value and the GAM model, we can conclude the correlation between EPM epigenetic age acceleration and sex is significant.

- **M1.3: Full Model with Sites Correlated with Chronological Age and Sex**

  M1.3 was also a full model. Since M1.2 did not perform as well as expected when including more sites correlated with different phenotype traits in the site selection process, we reduced the phenotype traits to only chronological age and sex. This was because in general, sites correlated with sex have higher PCC with methylation value. The 206 selected sites correlated with chronological age obtained in M1.1 was the same

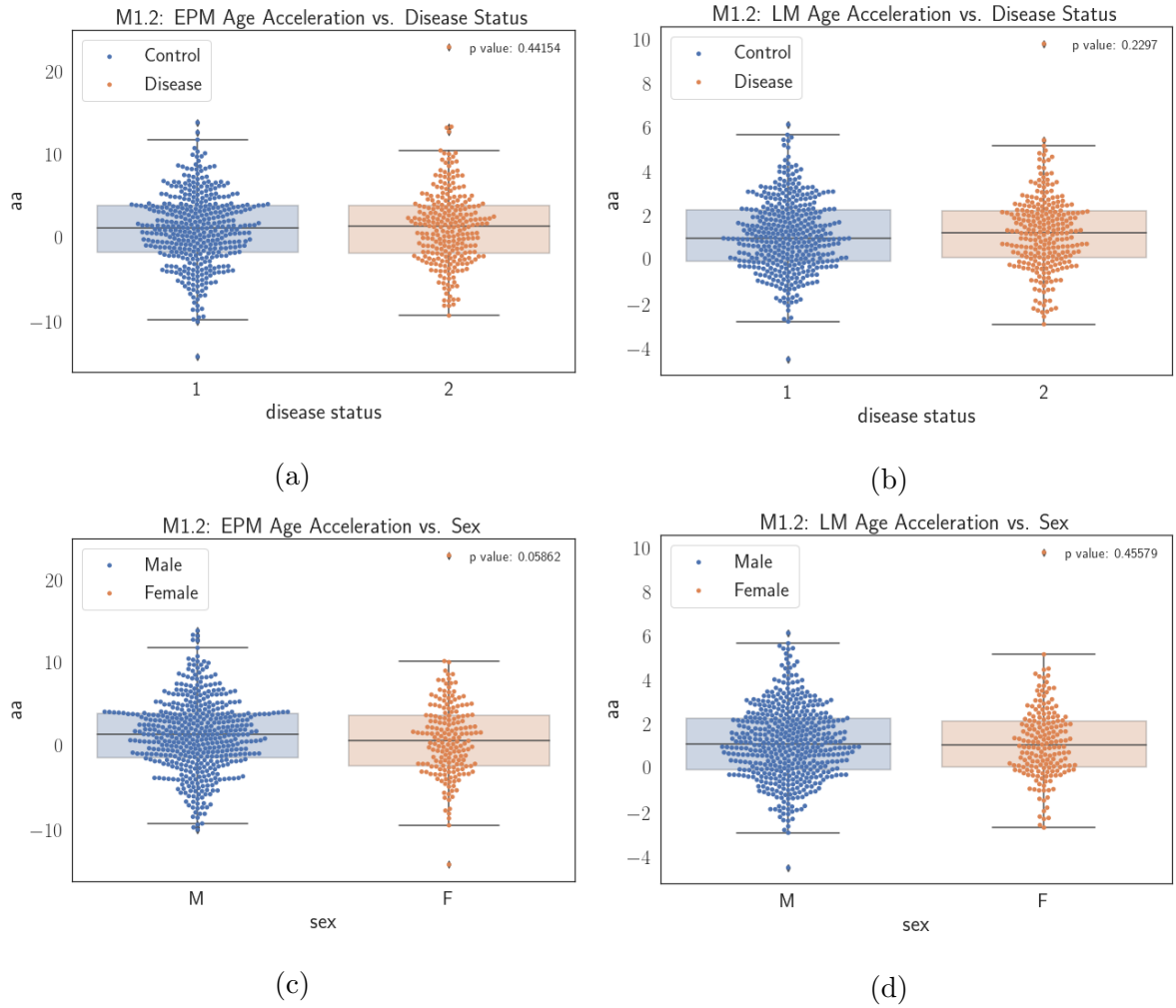


Figure 3.7: Age acceleration and phenotype traits of **M1.2:** Full Model with Sites Correlated with Chronological Age, Sex, and Disease.

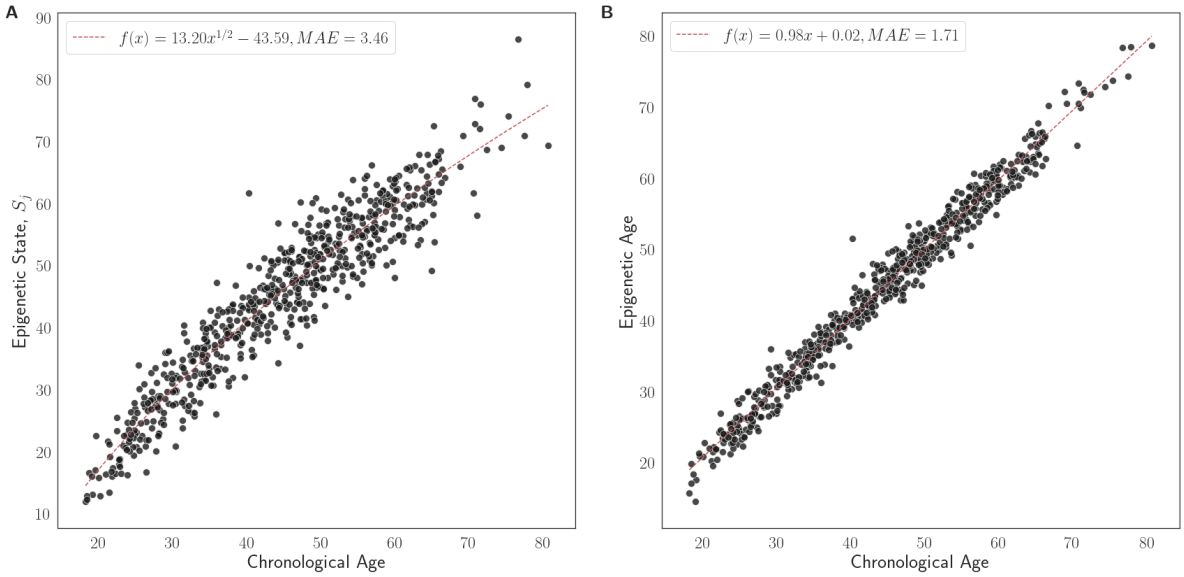
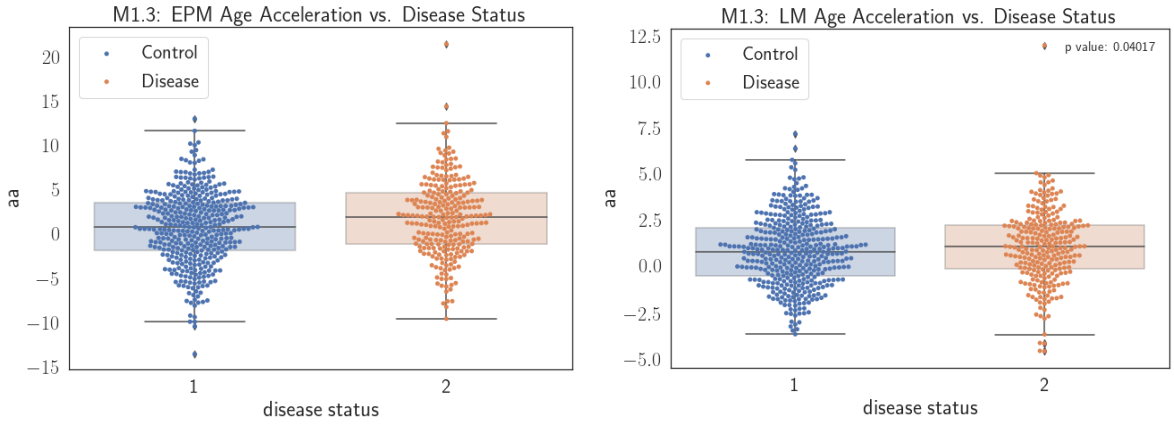


Figure 3.8: Chronological age and predicted epigenetic state of **A) M1.3:** Full Model with Sites Correlated with Chronological Age and Sex, and **B)** corresponding linear model

across M1.1, M1.2, and M1.3. The same PCC threshold of 0.58 was chosen to obtain sites correlated with sex, and there were 90 sites selected. There were no overlapping sites, and a total number of 296 sites were used to train the EPM model and the corresponding linear model. In 3.8, the corresponding linear model demonstrated a much tighter pattern around the predicted trend line of epigenetic state. The pattern of 3.8 was more similar to the pattern of 3.6 than 3.4.

The overall residuals of the linear model was significantly smaller than the EPM model. And in M1.3, not only the EPM model significantly separated phenotype traits with age acceleration, but the linear model also separated disease status with age acceleration. 3.9a and 3.9c revealed how well EPM age acceleration differentiates categories of disease status and sex respectively, with corresponding p-value of 0.00144 and 0.02852. 3.9b revealed how well linear model age acceleration differentiated categories of disease status, with a corresponding p-value of 0.04017. In general, samples with schizophrenia are having higher EPM age acceleration, and males are having higher EPM age



(a) (b)

(c) (d)

Figure 3.9: Age acceleration and phenotype traits of **M1.3:** Full Model with Sites Correlated with Chronological Age and Sex.

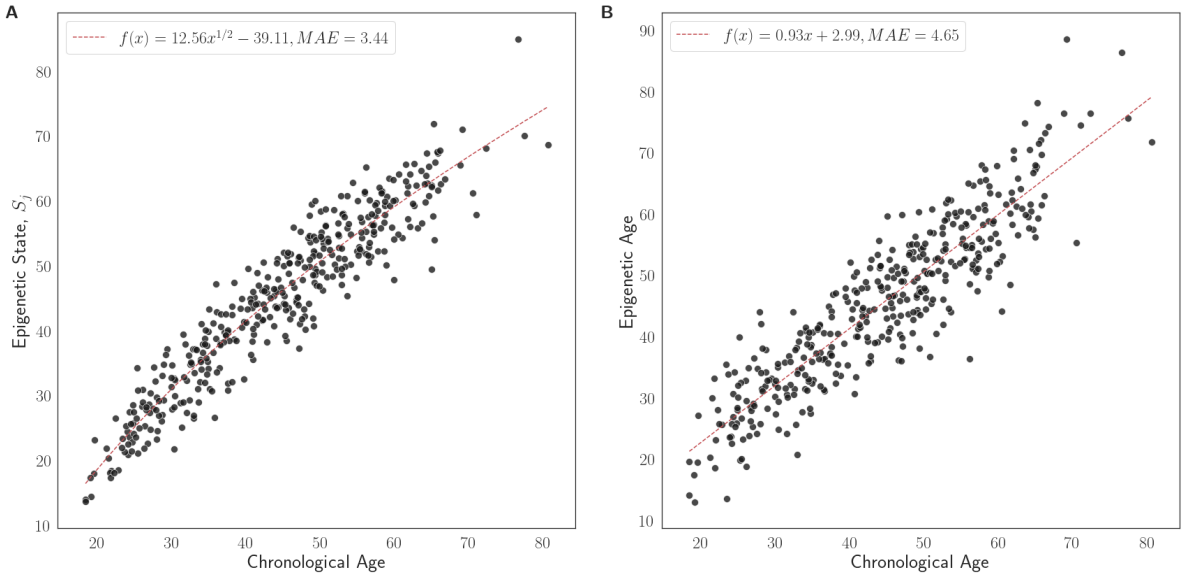


Figure 3.10: Chronological age and predicted epigenetic state of **A) M2.1:** CV TT Model with Sites Correlated with Chronological Age, and **B)** corresponding linear model

acceleration than females.

The results were consistent with the GAM model built on chronological age with epigenetic state, sex, and disease status. In the GAM model using EPM epigenetic state predictions, the parametric coefficients of chronological age, intercept, and disease status were significant. The covariate sex was slightly significant. While in GAM model using linear model epigenetic state predictions, only the parametric coefficients of chronological age and intercept were significant, and the parametric coefficient of disease status was slightly significant. In the EPM GAM model, both sex as a male and disease status as positive are having positive effects on epigenetic state. In the GAM model using linear age acceleration prediction, disease status has a positive effect on epigenetic state.

- **M2.1: CV TT Model with Sites Correlated with Chronological Age**

  For a CV TT, only methylation data of the training set were included when selecting sites and fitting the model. A training data set of 40% of all the samples were randomly

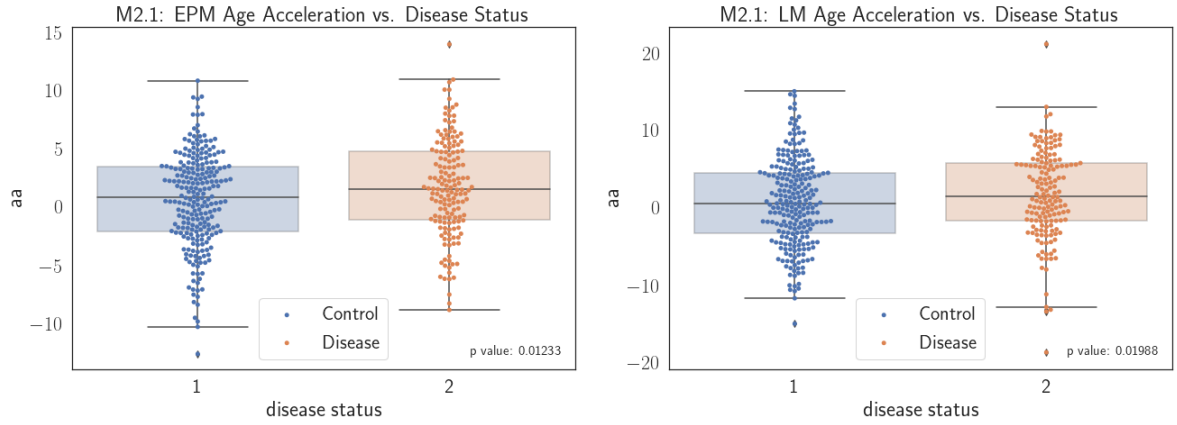
selected with stratification in chronological age, yielding 266 training samples. The remaining 399 samples were in the test set. A PCC threshold of 0.5936 was chosen to obtain 206 selected sites correlated with chronological age, keeping the number of sites selected the same as in the full model. In 3.10, the EPM model demonstrated a tighter pattern around the predicted trend line of epigenetic state.

Different from the full model, the overall residuals of the EPM model were now smaller than the linear model on the test data set. However, only disease was correlated with age acceleration from the EPM model, and age acceleration from the linear model. 3.11a and 3.11b revealed how well EPM age acceleration and linear model age acceleration differentiated categories of disease status respectively, with corresponding p-value of 0.01233 and 0.01988. In general, samples with schizophrenia are having higher EPM age acceleration.

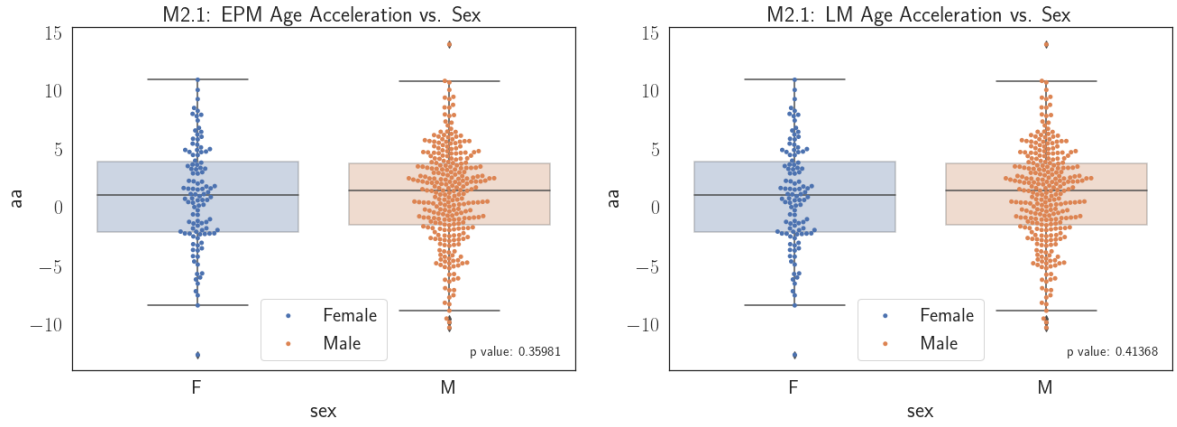
The results were consistent with the GAM model built on chronological age with epigenetic state, sex, and disease status. In the GAM model using EPM epigenetic state predictions, the parametric coefficients of chronological age, intercept, and disease status were significant. While in GAM model using linear model epigenetic state predictions, only the parametric coefficients of chronological age and intercept were significant. The disease covariate was slightly significant. Disease status as positive has a positive effect on epigenetic state.

- **M2.2: CV TT Model with Sites Correlated with Chronological Age, Sex, and Disease**

  M2.2 was also a CV TT model. The 206 selected sites correlated with chronological age obtained in M2.1 was the same across M2.1, M2.2, and M2.3. A PCC threshold of 0.565 was chosen to obtain 90 sites correlated with sex. To keep the same number of sites correlated with disease status and sites correlated with sex, a PCC threshold of 0.3975 was chosen to obtain 91 sites correlated with disease status. There were no



(a) (b)



(c) (d)

Figure 3.11: Age acceleration and phenotype traits of **M2.1:** CV TT Model with Sites Correlated with Chronological Age.

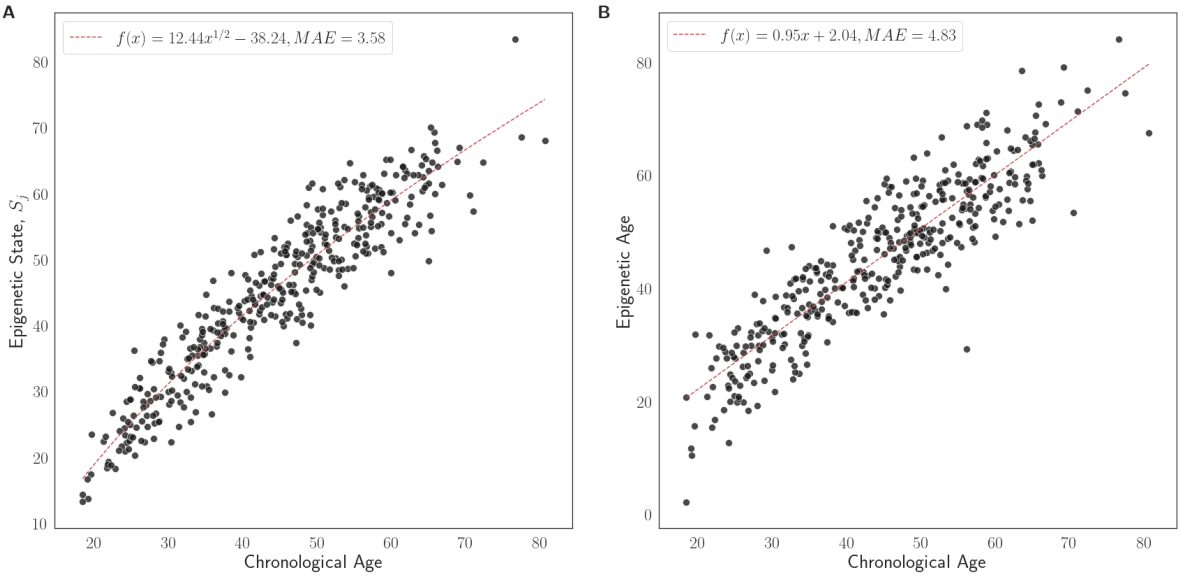


Figure 3.12: Chronological age and predicted epigenetic state of **A) M2.2:** CV TT Model with Sites Correlated with Chronological Age, Sex, and Disease, and **B)** corresponding linear model

overlapping sites, and a total number of 387 sites were used to train the EPM model and the corresponding linear model. In 3.12, the EPM model demonstrated a tighter pattern around the predicted trend line of epigenetic state.

Although the overall pattern of trend lines of the EPM model and the corresponding linear model in 3.12 were not so different from the ones in 3.10, the correlation between either EPM age acceleration or linear model age acceleration and phenotype traits were lower, as shown in 3.13. Only the EPM age acceleration and sex combination obtained a significant p-value. However, females are having higher age acceleration than males in this CV TT model, which was not consistent with previous findings.

The results were relatively consistent with the GAM model built on chronological age with epigenetic state, sex, and disease status. In both the GAM model using EPM epigenetic state predictions and the GAM model using linear predictions, only the parametric coefficients of chronological age and intercept were significant. GAM

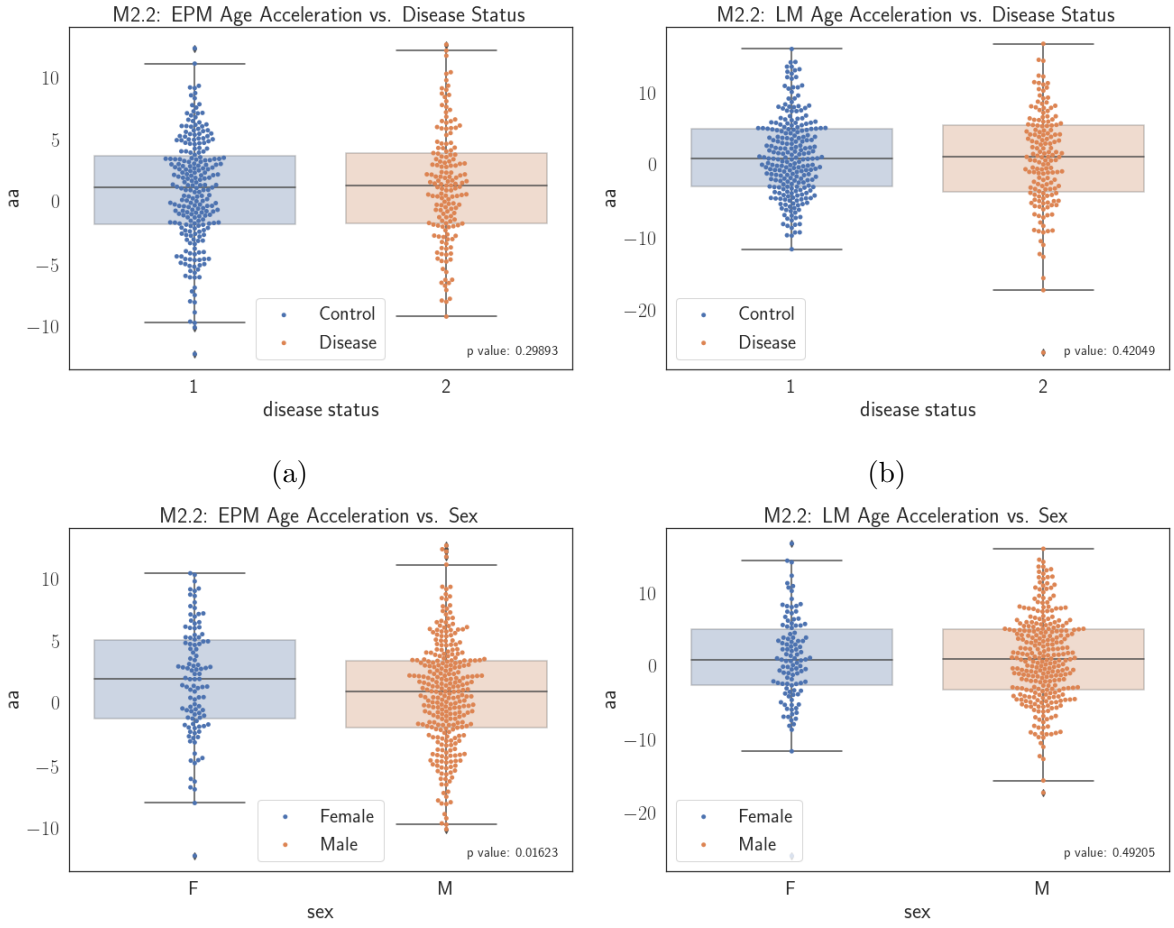


Figure 3.13: Age acceleration and phenotype traits of **M2.2:** CV TT Model with Sites Correlated with Chronological Age, Sex, and Disease.

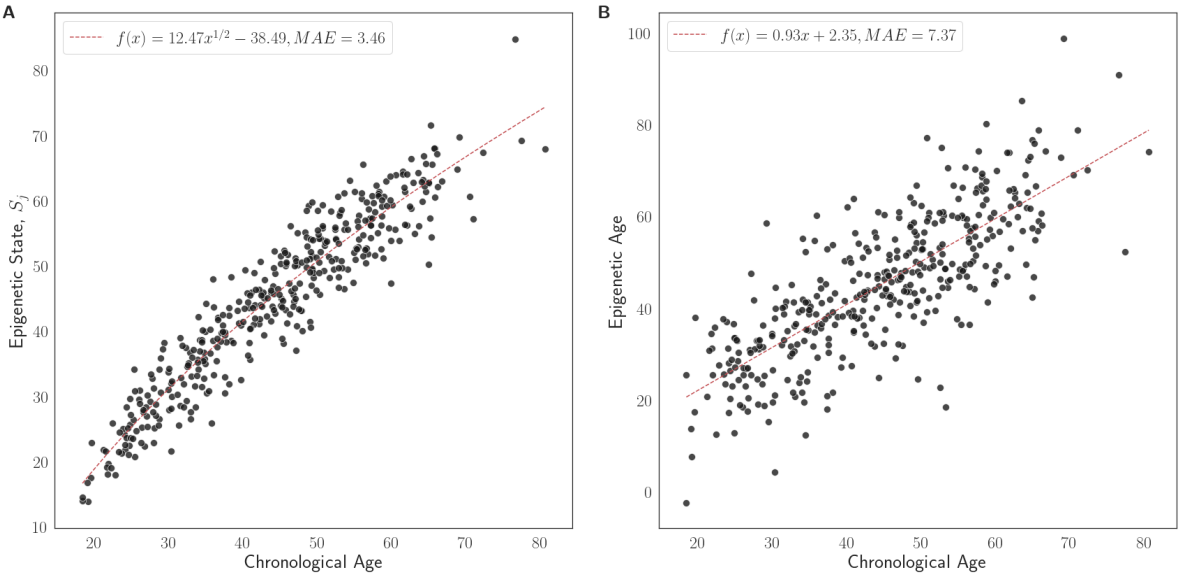


Figure 3.14: Chronological age and predicted epigenetic state of **A) M2.3:** CV TT Full Model with Sites Correlated with Chronological Age and Sex, and **B)** corresponding linear model

models suggested that sex and disease status were not sufficient to explain the residuals after chronological age was regressed out from epigenetic state prediction. There was an inconsistency between the direction of sex covariate, as it was positive as a male in the GAM model, while when directly correlating EPM age acceleration with sex, females are having higher age acceleration than male in general.

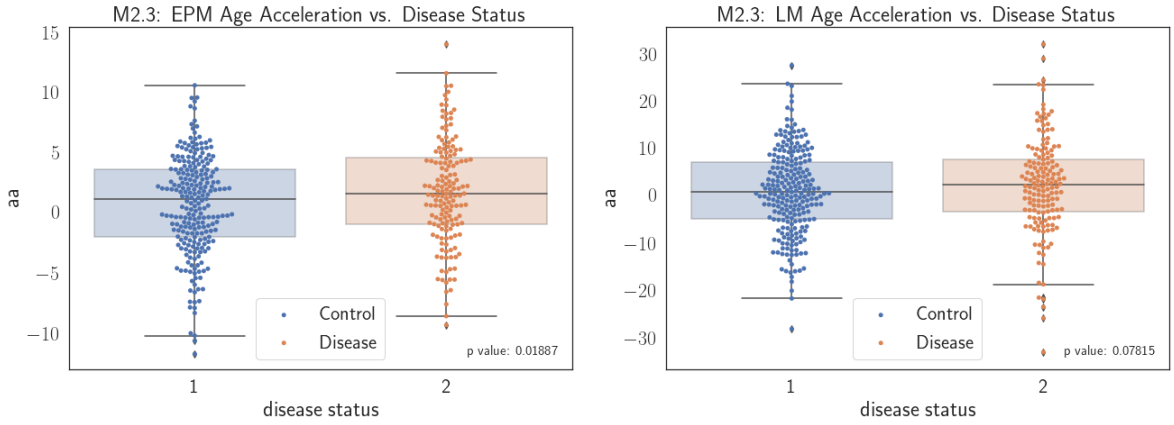
- **M2.3: CV TT Model with Sites Correlated with Chronological Age and Sex**

  Similarly, since M2.2 did not perform as well as expected when including more sites correlated with different phenotype traits in the site selection process, we reduced the phenotype traits to only chronological age and sex. This was because in general, sites correlated with sex have higher PCC with methylation value than sites correlated with disease status. The 206 selected sites correlated with chronological age obtained in M2.1 was the same across M2.1, M2.2, and M2.3. A PCC threshold of 0.565 was

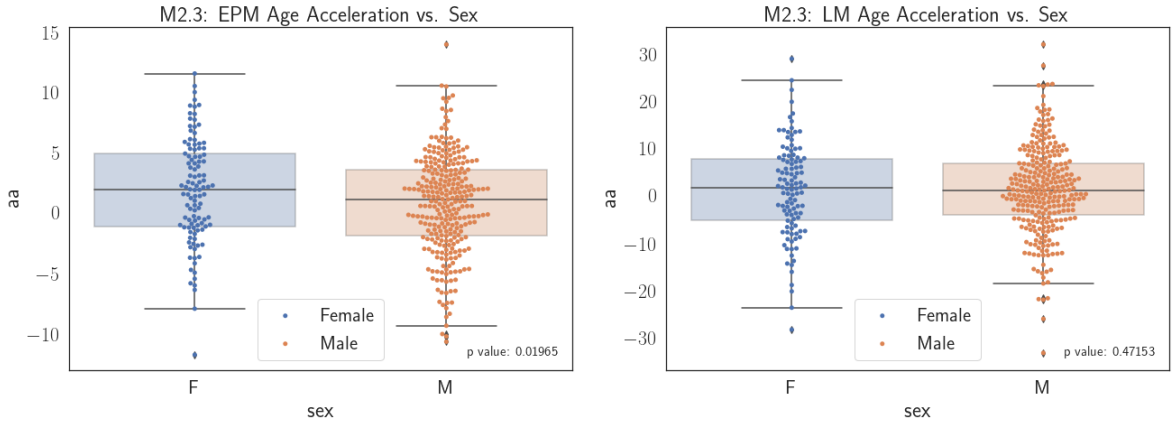
chosen to obtain 90 sites correlated with sex. There were no overlapping sites, and a total number of 296 sites were used to train the EPM model and the corresponding linear model. In 3.14, the EPM model demonstrated a much tighter pattern around the predicted trend line of epigenetic state. The linear model pattern in 3.14 was much sparser than the pattern of 3.10 and 3.12.

The overall residuals of the EPM model was significantly smaller than the conventional linear model. Only the EPM model significantly separated phenotype traits with age acceleration, while the linear model did not. 3.15a and 3.15c revealed how well EPM age acceleration differentiated categories of disease status and sex respectively, with corresponding p-value of 0.01887 and 0.01965. In general, samples with schizophrenia are having higher EPM age acceleration. However, in this training-testing split, females are having higher EPM age acceleration than males.

In terms of significance and directions, the results were consistent with the GAM model built on chronological age with epigenetic state, sex, and disease status. In the GAM model using EPM epigenetic state predictions, the parametric coefficients of chronological age, intercept, sex, and disease status were significant. While in GAM model using linear model epigenetic state predictions, only the parametric coefficients of chronological age and intercept were significant. The direction of sex as a male was negative with age acceleration in the EPM GAM model, which was not consistent with previous findings in the full model. This might attribute to the random sub-setting of the training and testing data set, and inadequate unbalanced sample numbers of male and female.



(a) (b)



(c) (d)

Figure 3.15: Age acceleration and phenotype traits of **M2.3:** CV TT Model with Sites Correlated with Chronological Age and Sex.

	EPM AA vs. Disease	LM AA vs. Disease	EPM AA vs. Sex	LM AA vs. Sex
M1.1	**0.00003**	0.22508	**0.02879**	0.02879
M1.2	0.44154	0.2297	**0.05862**	0.45579
M1.3	**0.00144**	**0.04017**	**0.02852**	0.45079
M2.1	**0.01233**	**0.01988**	0.35981	0.41368
M2.2	0.29893	0.42049	**0.01623**	0.49205
M2.3	**0.01887**	0.07815	**0.01965**	0.47153

Table 3.1: Significance Levels of Correlation between Various Model Predictions and Phenotype Traits

3.2.2 EPM models Can Capture the Relationship between Age Acceleration and Phenotype Traits

By examining the results from the previous section, we conclude that EPM models can differentiate individuals by specific phenotype traits using age acceleration prediction, when the EPM models are built appropriately with informative sites.

A further comparison of the EPM and linear model (LM) was carried out to show that the EPM model is capturing real biological differences than LM, which is only minimizing residuals statistically. Table 3.1 summarizes the p-values of correlating age acceleration (AA) from EPM and LM with phenotype traits, across all proposed models ranging from M1.1 to M2.3. In the table, significant p-values were bold. The p-value of 0.05862 was bold because we consider this term as significant combining evaluation results by both methods. By observing the EPM AA (Column 1 and Column 3) and the LM AA (Column 2 and Column 4), we can conclude that EPM models capture more biological information than conventional linear models trained with the same set of selected sites. Among different proposed models, there were many instances where EPM AA differentiated categories of phenotype traits while LM did not (M1.1 disease trait, M1.1 sex trait, M1.2 sex trait, M1.3 sex trait, M2.2 sex trait,

M2.3 disease trait, M2.3 sex trait). If LM can significantly differentiate a specific trait, the EPM is always also being able to do so (M1.3 disease trait and M2.1 disease trait).

3.2.3 A Full EPM Model Generates More Stable Epigenetic State Predictions

A set of full EPM models (M1.1, M1.2, M1.3) and a set of CV Train-Test Split models (M2.1, M2.2, M2.3) were compared. In terms of significance when differentiating phenotype traits by age acceleration, there was no big discrepancy among the two sets of models. However, in terms of consistency and stability, the full EPM model performed better than the CV TT model. The direction of how disease status is affecting age acceleration was consistent in all models. However, the direction of sex was not. In the full models, all models using different evaluation methods gave the same conclusion that males are generally having higher age acceleration than females, which is consistent with Horvath's findings on age acceleration of brain tissues using Horvath epigenetic clock [8]. In CV TT models, M2.2 and M2.3 revealed significance with sex and age acceleration. Both cases are suggesting females are having higher age acceleration in **??** and 3.15c. Although M2.1 did not show significance between sex and age acceleration, 3.11c and 3.11d revealed that males are having a slightly higher age accelerations in the box-plot. The results from the CV TT models were not consistent. As a result, a full EPM model is preferred when exploring the relationship between age acceleration and phenotype traits as it is more stable than a CV TT model.

3.2.4 Building EPM Model Using More Sites with Higher PCC Values Correlated with Phenotype Traits Is More Informative

Different sets of selected sites were experimented. M1.1 and M2.1 only utilized sites correlated with chronological age and yielded stable output when differentiating schizophrenia patients and controls. As M1.2 and M2.2 utilized more sites, including sites correlated with age, sex, and disease, the performance of the model decreased rather than increased. This

might be attributed to the fact that the new sites included in M1.2 and M2.2 were not all informative, increasing noise in the models. While the added sites correlated with sex have as high PCC (around 0.6) as sites correlated with age, the added sites correlated with disease status only had a threshold of 0.3. As a result, these sites with low PCC values were removed in further trials. M1.3 and M2.3 were built using only sites correlated with age and sex, which all have relatively high PCC values, and they performed significantly better than M1.2 and M2.2. As a result, we conclude that building an EPM model using sites with higher PCC values correlated with phenotype traits can achieve higher performance in capturing biological information.

In terms of the number of sites used, we further compared models with sites correlated with age, and models with sites correlated with both age and sex. For EPM full models, both M1.1 and M1.3 can capture the relationship between age acceleration and sex, disease. For CV TT EPM models, M2.3 was significant in terms of the relationship between age acceleration and sex, disease, while M2.1 was only significant in terms of disease. As a result, models with sites correlated with both age and sex were more informative when looking at both full models and CV TT models. This suggested that building models with more sites with high PCC values might help boost model performance.

CHAPTER 4

Discussion

Site selection methods are necessary when fitting an Epigenetic Pacemaker (EPM) clock as the Conditional Expectation Maximization algorithm used is computationally intensive. In the preprocessing step, Pearson Correlation Coefficient (PCC) method exhibits higher performance than the Randomized Ridge Regression method when fitting the EPM model once for different numbers of selected starting sites. In the model refitting step, site removal method by threshold could improve the performance of sites selected by Randomized Ridge Regression, but it is still no better than sites selected by PCC. The site removal method is unnecessary for sites selected by PCC. Overall, selecting sites by the PCC method in the preprocessing step and fitting an EPM model with those sites is the best site selection method for the EPM clock in this study.

The Randomized Ridge Regression method for site selection in the preprocessing step was expected to be faster than the PCC method. However, it showed that there is no significant difference between the runtime of the Randomized Ridge Regression method and the PCC method. Moreover, the PCC method generated the optimal EPM model with 450 starting sites, which is about one-third of the starting sites (1565 sites) needed for the Randomized Ridge Regression method to generate its optimal EPM model. This shows that PCC method not only achieved a higher model performance, but also developed an EPM model with less complexity.

As a result, PCC site selection method was further used when applying the EPM model to real data sets. By experimenting with different proposed EPM models with options in

building model and site selection, we conclude that EPM models can differentiate individuals by specific phenotype traits using age acceleration prediction when built appropriately with informative sites. How to choose informative sites is always of great importance to building a good EPM model, as the predictions strongly rely on the input methylation values of specific sets. By trying different sets of selected sites and comparing the models by two different evaluation methods, we can conclude that sites with higher PCC values are informative sites. In this case, a threshold of 0.58 is a feasible cutoff. Also, when the number of sites used is still at the scale of hundreds, we speculate that the more the sites with high PCC values are used, the better the EPM model will perform.

Intuitively, using a full model will perform better than a CV TT model, because there are more samples included when fitting the model. Since the EPM model optimized the residuals of the methylation values on each site, the predicted epigenetic state will not be over-fitted on chronological ages. Furthermore, as it was found that the predictions from a full EPM model and CV full EPM model yielded similar results (the coefficients of trend line function only differed by 0.01), it is reasonable to conclude that over-fitting is a trivial problem for the EMP model. Also, the goal is to find whether there are correlations between epigenetic age acceleration and phenotype traits, and the results are more meaningful for exploratory purposes than predictive purposes. As a result, it is sufficient to conclude that using a full EPM model is the best choice in this case.

However, the usage of GAM function on a small number of samples in the evaluation process still requires further study. It was noticed that in EPM model of M2.2, the distribution was clear that females are having higher age acceleration than males, as shown in **??**. In the GAM function, it was shown that sex as a male is having a positive effect, contradicting the distribution in the boxplot. This inconsistency might be due to the imbalanced samples of female and male in the test data set, and the sample size in the test data set might be too small.

Future directions include obtaining methylation data with more continuous phenotype

data, such as BMI, etc, then there will be more traits that we can correlate age acceleration with. Also, modeling chronological as a function of predicted epigenetic state and other phenotype traits would be more accurate. We may also test the generalization ability of the EPM clock, applying an EPM clock built on one data set to another data set. We can also compare the EPM clock with other built epigenetic clocks in terms of the ability to separate patients and controls.

CHAPTER 5

Supplement Figures

5.1 Summary Tables of GAM Evaluation Method

5.2 Fitting Curves of GAM Evaluation Method

```
Family: gaussian
Link function: identity

Formula:
epm_pred ~ Sex + disease_status + s(age)

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      43.6045     0.3244 134.402  < 2e-16 ***
SexM              0.6191     0.3435   1.802    0.072 .
disease_status2   1.4207     0.3201   4.438 1.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(age) 3.354  4.224 1702  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.915   Deviance explained = 91.6%
GCV = 15.764  Scale est. = 15.614    n = 665
```

(a) EPM prediction

```
Family: gaussian
Link function: identity

Formula:
lm_pred ~ Sex + disease_status + s(age)

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      44.3985     0.1773 250.443   <2e-16 ***
SexM              0.2200     0.1877   1.172    0.242
disease_status2   0.1267     0.1749   0.724    0.469
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(age) 3.257  4.107 5476  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.971   Deviance explained = 97.2%
GCV =  4.708  Scale est. = 4.6637    n = 665
```

(b) LM prediction

Figure 5.1: GAM Summary Table of **M1.1:** Full Model with Sites Correlated with Chronological Age

```
Family: gaussian
Link function: identity

Formula:
epm_pred ~ Sex + disease_status + s(age)

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      43.5374     0.3388 128.510  < 2e-16 ***
SexM              1.2538     0.3587   3.495 0.000505 ***
disease_status2   0.4205     0.3343   1.258 0.208882
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(age) 3.367   4.24 1548  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.908   Deviance explained = 90.9%
GCV =  17.19  Scale est. = 17.025    n = 665
```

(a) EPM prediction

```
Family: gaussian
Link function: identity

Formula:
lm_pred ~ Sex + disease_status + s(age)

Parametric coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     44.623509   0.137565 324.382   <2e-16 ***
SexM            -0.025932   0.145663  -0.178    0.859
disease_status2  0.005052   0.135716   0.037    0.970
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(age) 3.292  4.149 9220  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.983   Deviance explained = 98.3%
GCV = 2.8347  Scale est. = 2.8078    n = 665
```

(b) LM prediction

Figure 5.2: GAM Summary Table of **M1.2:** Full Model with Sites Correlated with Chronological Age, Sex, and Disease

```
Family: gaussian
Link function: identity

Formula:
epm_pred ~ Sex + disease_status + s(age)

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      43.7093     0.3432 127.368  < 2e-16 ***
SexM              0.6046     0.3633   1.664 0.096553 .
disease_status2   1.1792     0.3387   3.482 0.000531 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(age) 3.394  4.273 1504  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.906   Deviance explained = 90.7%
GCV = 17.637  Scale est. = 17.467    n = 665
```

(a) EPM prediction

```
Family: gaussian
Link function: identity

Formula:
lm_pred ~ Sex + disease_status + s(age)

Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     44.51761    0.15520 286.835   <2e-16 ***
SexM            -0.01753    0.16443  -0.107   0.9151
disease_status2  0.26041    0.15287   1.703   0.0889 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df    F p-value
s(age) 2.874  3.637 8173  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.978   Deviance explained = 97.8%
GCV =  3.612  Scale est. = 3.5801    n = 665
```

(b) LM prediction

Figure 5.3: GAM Summary Table of **M1.3:** Full Model with Sites Correlated with Chronological Age and Sex

```
Family: gaussian
Link function: identity

Formula:
epm_pred ~ sex + disease + s(age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.1552     0.4415 100.016  < 2e-16 ***
sexM          0.2212     0.4658   0.475  0.63517
disease2      1.1606     0.4319   2.687  0.00752 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(age) 2.856  3.623 977.1  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.899   Deviance explained =   90%
GCV = 17.384  Scale est. = 17.129    n = 399
```

(a) EPM prediction

```
Family: gaussian
Link function: identity

Formula:
lm_pred ~ sex + disease + s(age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.6480     0.6068  73.575   <2e-16 ***
sexM          0.4937     0.6425   0.769   0.4426
disease2      1.0521     0.5859   1.796   0.0733 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
       edf Ref.df    F p-value
s(age)   1      1 1764  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.816   Deviance explained = 81.7%
GCV = 33.043  Scale est. = 32.712    n = 399
```

(b) LM prediction

Figure 5.4: GAM Summary Table of **M2.1:** CV TT Model with Sites Correlated with Chronological Age

```
Family: gaussian
Link function: identity

Formula:
epm_pred ~ sex + disease + s(age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.4810     0.4451  99.929   <2e-16 ***
sexM          0.2306     0.4696   0.491    0.624
disease2      0.5095     0.4355   1.170    0.243
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(age) 2.861  3.629 984.4  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =    0.9   Deviance explained = 90.1%
GCV = 17.672  Scale est. = 17.412    n = 399
```

(a) EPM prediction

```
Family: gaussian
Link function: identity

Formula:
lm_pred ~ sex + disease + s(age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.8565     0.5539  80.986   <2e-16 ***
sexM          0.6979     0.5842   1.195    0.233
disease2      0.1827     0.5422   0.337    0.736
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(age) 3.041  3.851 574.7  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.848   Deviance explained =   85%
GCV = 27.347  Scale est. = 26.933    n = 399
```

(b) LM prediction

Figure 5.5: GAM Summary Table of **M2.2:** CV TT Model with Sites Correlated with Chronological Age, Sex, and Disease

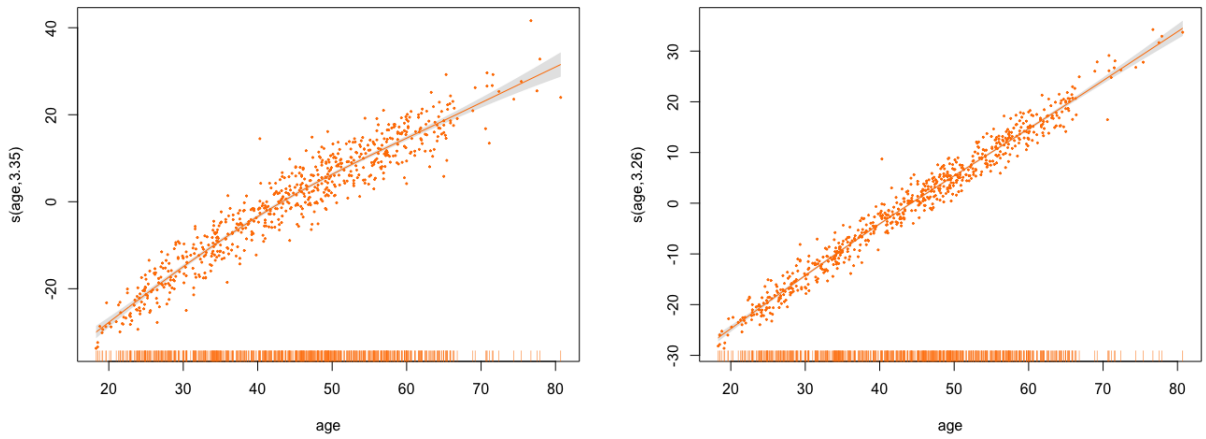
```
Family: gaussian
Link function: identity

Formula:
epm_pred ~ sex + disease + s(age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  45.0914     0.4432 101.737   <2e-16 ***
sexM         -0.9598     0.4676  -2.053   0.0408 *
disease2      1.0758     0.4337   2.481   0.0135 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
         edf Ref.df     F p-value
s(age) 2.893  3.668 946.8  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.897   Deviance explained = 89.8%
GCV = 17.519  Scale est. = 17.26     n = 399
```

(a) EPM prediction

```
Family: gaussian
Link function: identity

Formula:
lm_pred ~ sex + disease + s(age)

Parametric coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  44.2757     1.0136  43.683   <2e-16 ***
sexM          0.1441     1.0731   0.134    0.893
disease2      1.3442     0.9785   1.374    0.170
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Approximate significance of smooth terms:
       edf Ref.df     F p-value
s(age)   1      1 637.5  <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.615   Deviance explained = 61.8%
GCV = 92.181  Scale est. = 91.257    n = 399
```

(b) LM prediction

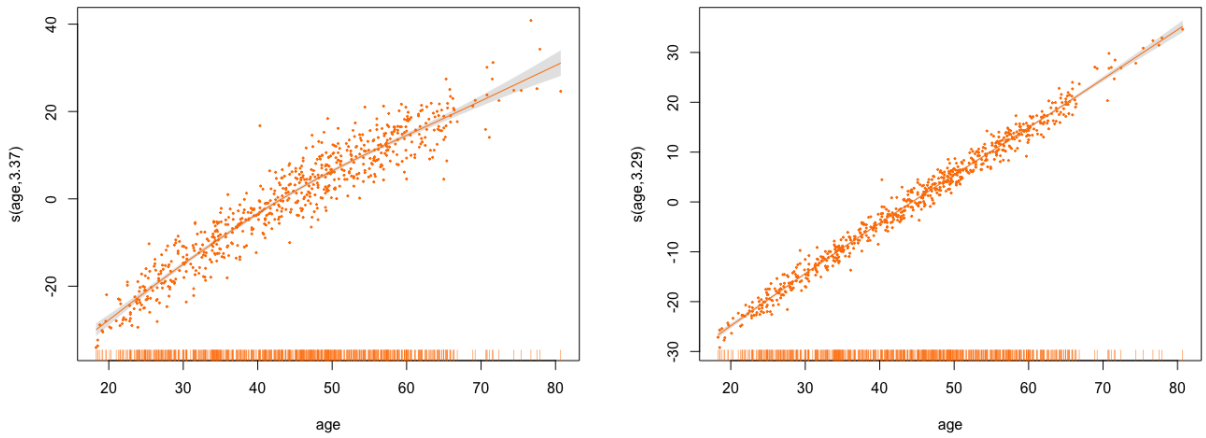
Figure 5.6: GAM Summary Table of **M2.3:** CV TT Model with Sites Correlated with Chronological Age and Sex



(a) EPM prediction

(b) LM prediction

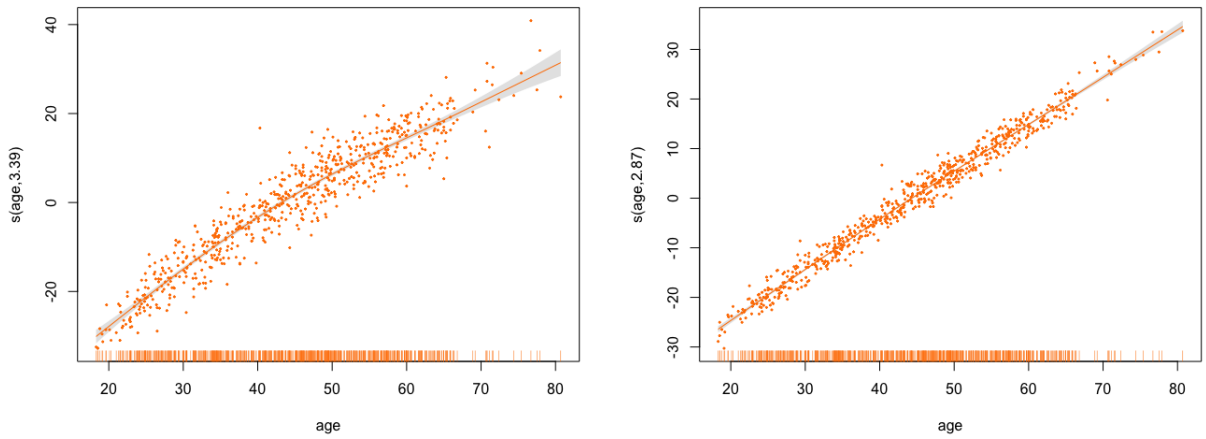
Figure 5.7: GAM Fitting Curve of **M1.1:** Full Model with Sites Correlated with Chronological Age



(a) EPM prediction

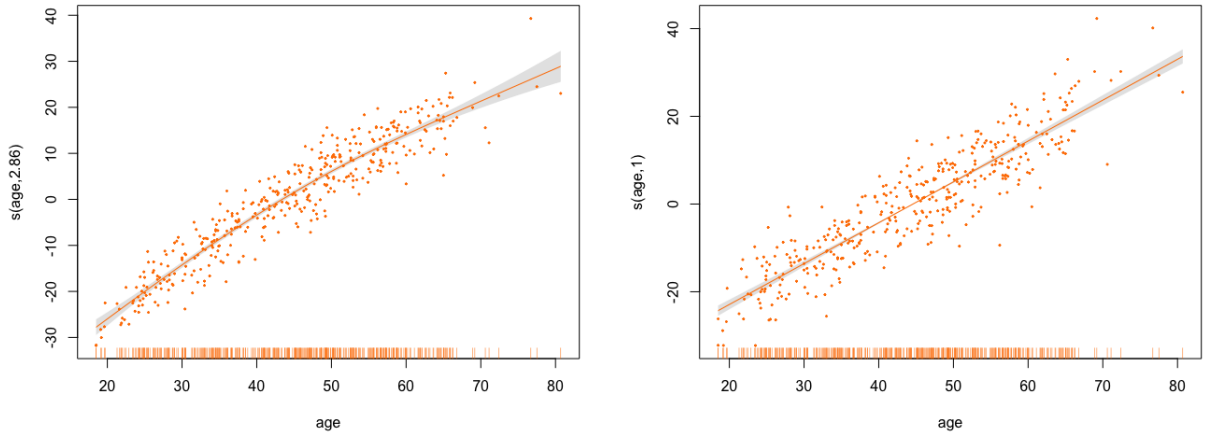
(b) LM prediction

Figure 5.8: GAM Fitting Curve of **M1.2:** Full Model with Sites Correlated with Chronological Age, Sex, and Disease



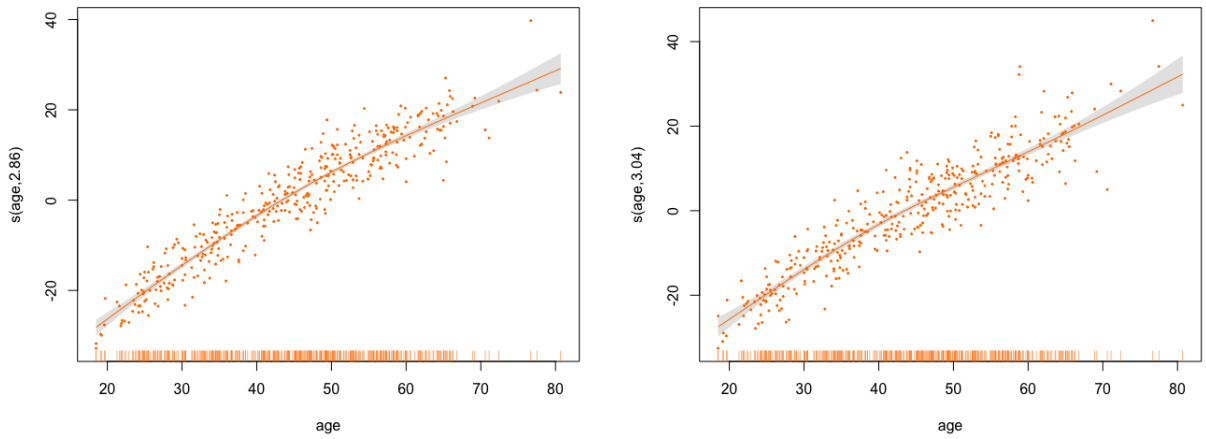
(a) EPM prediction (b) LM prediction

Figure 5.9: GAM Fitting Curve of **M1.3:** Full Model with Sites Correlated with Chronological Age and Sex



(a) EPM prediction (b) LM prediction

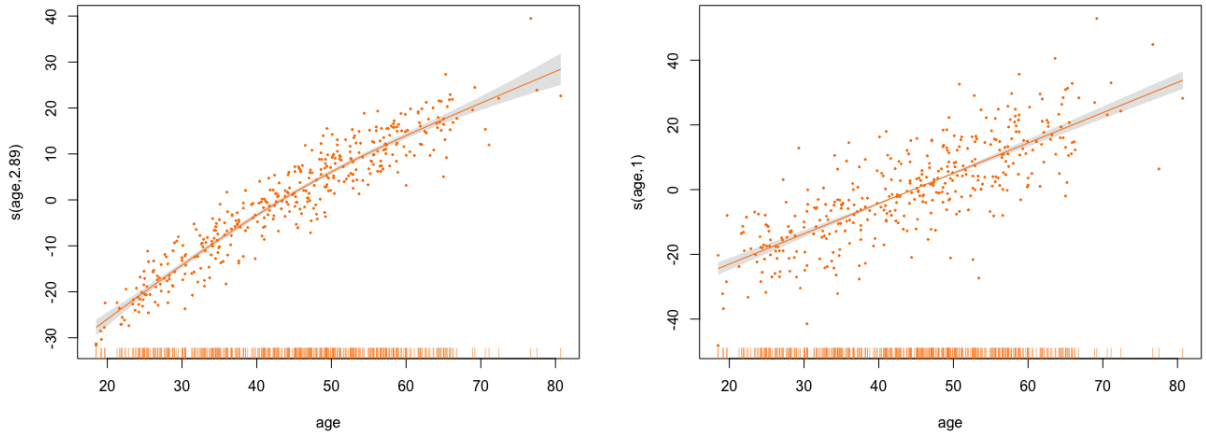
Figure 5.10: GAM Fitting Curve of **M2.1:** CV TT Model with Sites Correlated with Chronological Age



(a) EPM prediction

(b) LM prediction

Figure 5.11: GAM Fitting Curve of **M2.2:** CV TT Model with Sites Correlated with Chronological Age, Sex, and Disease



(a) EPM prediction

(b) LM prediction

Figure 5.12: GAM Fitting Curve of **M2.3:** CV TT Model with Sites Correlated with Chronological Age and Sex

REFERENCES

[1] Steve Horvath. DNA methylation age of human tissues and cell types. *Genome Biology*, 14(10):3156, December 2013.

[2] Assaf Zemach, Ivy E. McDaniel, Pedro Silva, and Daniel Zilberman. Genome-wide evolutionary analysis of eukaryotic dna methylation. *Science*, 328(5980):916–919, 2010.

[3] George T. Baker and Richard L. Sprott. Biomarkers of aging. *Experimental Gerontology*, 23(4):223 – 239, 1988.

[4] Steve Horvath, Paolo Garagnani, Maria Giulia Bacalini, Chiara Pirazzini, Stefano Salvioli, Davide Gentilini, Anna Maria Di Blasio, Cristina Giuliani, Spencer Tung, Harry V. Vinters, and Claudio Franceschi. Accelerated epigenetic aging in down syndrome. *Aging Cell*, 14(3):491–495, 2015.

[5] Gregory Hannum, Justin Guinney, Ling Zhao, Li Zhang, Guy Hughes, SriniVas Sadda, Brandy Klotzle, Marina Bibikova, Jian-Bing Fan, Yuan Gao, Rob Deconde, Menzies Chen, Indika Rajapakse, Stephen Friend, Trey Ideker, and Kang Zhang. Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates. *Molecular Cell*, 49(2):359–367, January 2013.

[6] Steve Horvath and Kenneth Raj. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics*, 19(6):371–384, June 2018.

[7] Sagi Snir, Colin Farrell, and Matteo Pellegrini. Human epigenetic ageing is logarithmic with time across the entire lifespan. *Epigenetics*, 14(9):912–926, 2019. PMID: 31138013.

[8] Steve Horvath, Michael Gurven, Morgan E Levine, Benjamin C Trumble, Hillard Kaplan, Hooman Allayee, Beate R Ritz, Brian Chen, Ake T Lu, Tammy M Rickabaugh, Beth D Jamieson, Dianjianyi Sun, Shengxu Li, Wei Chen, Lluis Quintana-Murci, Maud Fagny, Michael S Kobor, Philip S Tsao, Alexander P Reiner, Kerstin L Edlefsen, Devin Absher, and Themistocles L Assimes. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome biology*, 17(1):171–171, August 2016.