Healthy Aging Signal Research
Milestone 3 Technical Report

Partner:
Merck & Co.
Group Members:
Eleonora Shantsila: eshantsila@g.harvard.edu,
Yaxin Lei: yaxin_lei@g.harvard.edu,
Aaron Jacobson: aaronjacobson@g.harvard.edu ,
Daniel Cox: daniel_cox@g.harvard.edu

1 . Problem description

Merck is one of the world's largest pharmaceutical companies, investing billions of dollars annually in research areas such as oncology, vaccines, infectious and neurodegenerative diseases. We have been working with data scientists from Merck who are interested in modeling the process of healthy aging, as a great many diseases develop more readily with age. For the purposes of this project, we define "unhealthy" as individuals with neurodegenerative diseases, for instance Huntington's, Parkinson's and Alzheimer's.

At the outset of this project we defined the following goals.

1. To identify databases that are relevant to aging;
2. To identify features that are good biomarkers of healthy aging;
3. To build age-predictive models based on the biomarkers we identify.
4. To compare the aging process between healthy and unhealthy cohorts, as well as biomarkers from different tissues.

We then set as our initial goal to develop a model that could predict age based on biomarkers to within a mean absolute error (MAE) of 5 years. We chose this value —perhaps naively— because we thought it may be close to the lower limit of accuracy of age-predictive models, given the biological variation observed in individuals of the same age. As detailed below, we have accomplished this goal. Indeed our best model has an MAE of less than 3.6 years. We hope this model may be used in the future as a baseline against which the aging process can be gauged in specific individuals and then the resulting information used to assess the need for therapeutic intervention to prevent the onset of disease.

2. Literature Overview

To begin we considered several types of biomarkers that had been identified in the biological literature as potentially being indicative of age (Langer et al, 2019; Tozer et al, 2018; Salameh et al., 2018; Putin et al, 2016, Frenk and Houseley, 2018). These included brain-scanned images, blood levels of various biological molecules, RNA expression levels, and DNA methylation patterns. We then explored these forms of data with data from the PPMI Parkinson's

database (https://www.ppmi-info.org), which contains data from both healthy and diseased individuals. This work led us to conclude that DNA-methylation data would be most useful, and at that point we focused our work in this area. Also we discovered that there was a good deal of literature that discusses DNA-methylation vis-a-vis aging.

DNA methylation refers to a phenomenon where methyl groups are attached to various sites in an organism's DNA over the course of its lifetime. This may occur differently in different tissues, and the methylation status of a cell can have important effects on gene expression. In the last several years a series of studies have been published that correlate DNA methylation with aging and present models to relate the two (for review see Salameh et al. 2020). Notably, Horvath in 2013 examined data from a wide variety of tissues and datasets and used an elastic-net linear model to identify 353 methylation sites (cpgs sites) out of many thousands whose methylation state is most predictive of age. Based on these 353 sites, he was able to predict age in test datasets to within a median value of less than 4 years. Similar results were also obtained by Hannum et al (2013). Working with 656 samples from whole blood, they also used an elastic-net model to identify 71 age-related methylation sites that they then used to create a predictive model with comparable accuracy to that of Horvath (2013). Further, a similar study was performed by Weidner et al in 2014. In their study 102 age-related methylation sites were identified also from whole blood samples. Perhaps surprisingly, however, there was very little overlap between the age-related methylation sites identified in each study. Hannum and Horvath shared only 6 common sites, and Hannum and Weidner and Weidner and Horvath only 1. Thus, overall, there is a good deal of interest in relating DNA methylation to aging and a growing literature, but as of yet investigators have not come to a consensus on the most relevant DNA methylation sites and how these might vary with tissue type. We decided to explore this area ourselves.

3. Data Overview

3.1 A New Source of DNA Methylation Data

Following the literature review two potential markers that could be used for chronological age prediction were selected, blood chemistry and methylation. EDA was conducted using both of these markers and the PPMI database to determine the most promising marker to conduct in depth analysis on. The blood chemistry EDA can be found in Appendix A. Due to the results indicating no clear relationship between chronological age and blood chemistry results in our data this avenue was not pursued further with the remainder of the work conducted using DNA methylation data.

DNA methylation EDA was conducted on the PPMI database which had around 200 patient samples. Due to the very large number of features associated with methylation data (up to 450,000) a large source of data was needed, namely the Epigenome Wide Association Studies (EWAS) database, which houses methylation data from many tissues and from both healthy and diseased individuals (https://bigd.big.ac.cn/ewas/datahub).

All further analysis in this report was conducted on this data with the goal of the exploration to answer the following questions:

- **Model accuracy**: How accurately can we predict the age of healthy individuals using DNA methylation data?
- **Feature importance**: What methylation sites are most important in predicting age?
- **Model generalization:** Do these sites differ from tissue to tissue?
- **Model generalization:** Do models trained on healthy individuals transfer to unhealthy individuals?

3. 2 EWAS Database overview

The EWAS database contains data on a vast variety of tissues, healthy and unhealthy cohorts as well as including metadata for some of the patients. In this analysis we focused on the whole blood, brain and breast tissue data only. Additionally, we used healthy control, Alzheimer's, Parkinson's and Huntington's subject cohorts for the analysis. Due to the dataset being a combination of data from a large number of different studies, metadata available varied from subject to subject. It consistently included age, which we used for our analysis, however, other metadata available for some subjects such as smoking status and HIV status were not explored in detail.

The table in Figure 1 shows a summary of the sample sizes across different tissues and cohorts relevant to the analysis.

	Healthy Control	Alzheimer's	Parkinson's	Huntington's
Whole Blood	1802	111	222	N/A
Brain	1064	811	N/A	270
Breast	520	N/A	N/A	N/A

Figure 1: Sample sizes of healthy and unhealthy cohorts in EWAS database across whole blood, brain and breast tissue

As our analysis is focused on predicting age, we require the ages of the samples to be well distributed (namely for a wide range of ages to be present with as close to even sampling as possible). Figure 2 A-C shows the distribution of ages in each of the healthy cohorts across the three tissues we are interested in. From the figure we can see that the distributions, particularly for the whole blood tissue data are fair even across age groups and cover a wide array of ages.

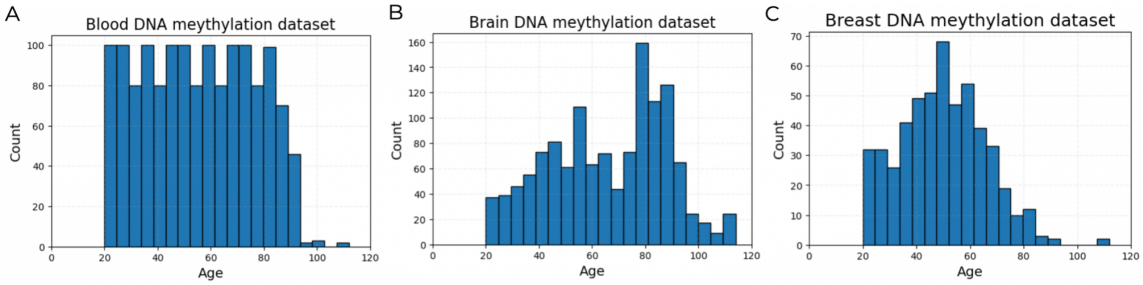


Figure 2: Histograms showing the distribution of ages in the A) whole blood, B) brain and C) breast tissues in the

healthy control cohorts.

Figure 3 shows the age distribution for the unhealthy cohorts. Although having well distributed ages (well distributed defined as above), is not necessary for the unhealthy cohorts, as they are used for comparison rather than training, these are still relevant to the model performance interpretations and are provided for completeness.

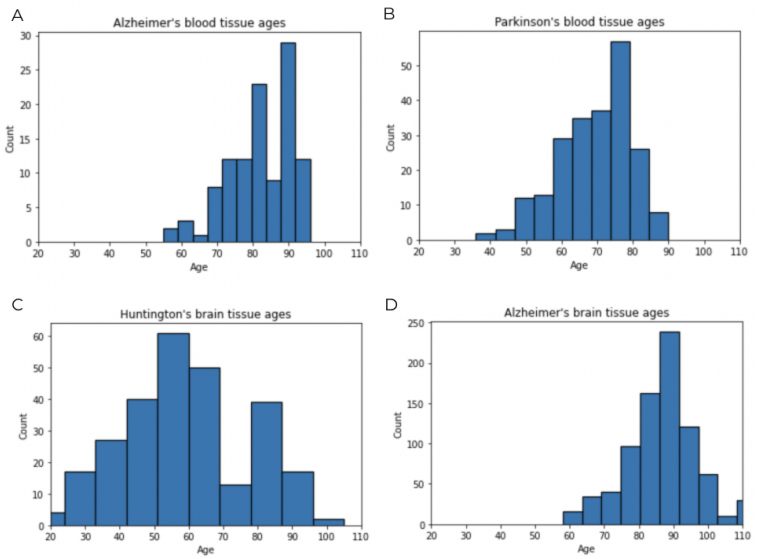


Figure 3: Histograms showing the distribution of ages in the A)-B) whole blood unhealthy cohorts and C)-D) brain tissue unhealthy cohorts.

After loading and processing, the methylation data from EWAS looks as shown in Fig. 4. Each row represents a sample from a different person. Each column —apart from the “age” and “tissue” columns— represents a potential methylation site (CpG site). Each value in the table indicates the probability that a particular cpg is methylated in that particular sample. Of note, this table contains over 375,603 columns, so the dataset is very large and contains an enormous number of features.

sample_id	tissue	age	cg02494853	cg03706273	cg04023335	cg05213048	cg15295597	cg26520468	cg27539833	cg00008
GSM2334366	whole blood	94	0.078	0.205	0.139	0.904	0.120	0.970	0.912	0.
GSM989863	whole blood	101	0.013	0.008	0.117	0.756	0.033	0.958	0.933	0.
GSM1443696	whole blood	99	0.013	0.017	0.477	0.715	0.017	0.966	0.932	0.
GSM1069241	whole blood	99	0.013	0.017	0.477	0.715	0.017	0.966	0.932	0.
GSM1572442	whole blood	112	0.036	0.255	0.260	0.690	0.065	0.983	0.951	0.
...	...	...	...	...	...	...	...	...	...	
GSM1498536	whole blood	48	0.010	0.048	0.068	0.575	0.034	0.981	0.946	0.
GSM1868331	whole blood	48	0.024	0.019	0.635	0.848	0.035	0.958	0.944	0.
GSM2337452	whole blood	48	0.027	0.032	0.145	0.661	0.068	0.964	0.936	0.
GSM1653326	whole blood	48	0.033	0.023	0.529	0.772	0.064	0.956	0.946	0.
GSM1871289	whole blood	48	0.019	0.024	0.166	0.599	0.048	0.952	0.949	0.

1066 rows × 375603 columns

Figure 4. DNA methylation data from the EWAS database for healthy individuals and whole blood tissue

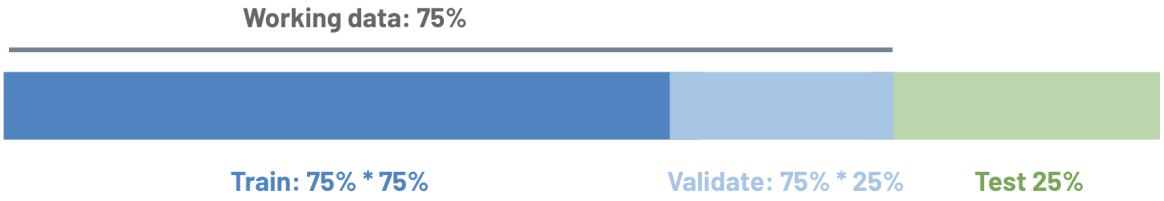
4. Age Prediction Modeling

4.1 Standard Data Handling Procedure

Due to the size and volume of data, and the fact that we explored multiple models for age prediction, we've came up with a standard procedure that is employed that has been tested to yield the best performance before all modeling work.

When dealing with NAs, we dropped all columns with greater than 10% of NAs before any splits. Other methodologies we tested included dropping 25% percent of NAs, and imputing NA entries with KNN. But they either yielded worse performance on test data, or yielded slight performance boost but too time consuming in the case of KNN. Since DNA methylation happens as people ages, we also removed young individuals who are under 20 years old which may have unpredicted methylation data.

We also standardized the ways that validation and test splits, as well as NA imputations are computed. For the remainder of our report, we use the following terminologies. Data is first split into working data (75%) and test (25%). By test we mean a held-out portion of data for final performance testing not involved in the training process. The test set NA is imputed by using the mean of working data. The working data is further divided into train set (75% * 75%) and validation set (75% * 25%) when necessary for parameter tuning and model refinement purposes, where the validate NA is imputed by using the train split mean.



4.2 Healthy Cohort Age Predictive Modeling

4.2.1 Age Prediction Modeling - Whole Blood Tissue:

After examining the various tissue types in the database, we started our work with methylation data from whole blood, as for healthy individuals this was most abundant. We began by asking how well we could predict age using all 375,603 features. To investigate this, we performed multiple linear regression with all the features and age as the dependent variable. The results are shown in Fig. 5. Encouragingly, with or without regularization, linear models predict age fairly well over the entire lifespan —with a MAE on a test dataset of ~ 4 years. We also used a nonlinear tree-based regression method on these data, XGboost and found some improvement (Fig. 5D).

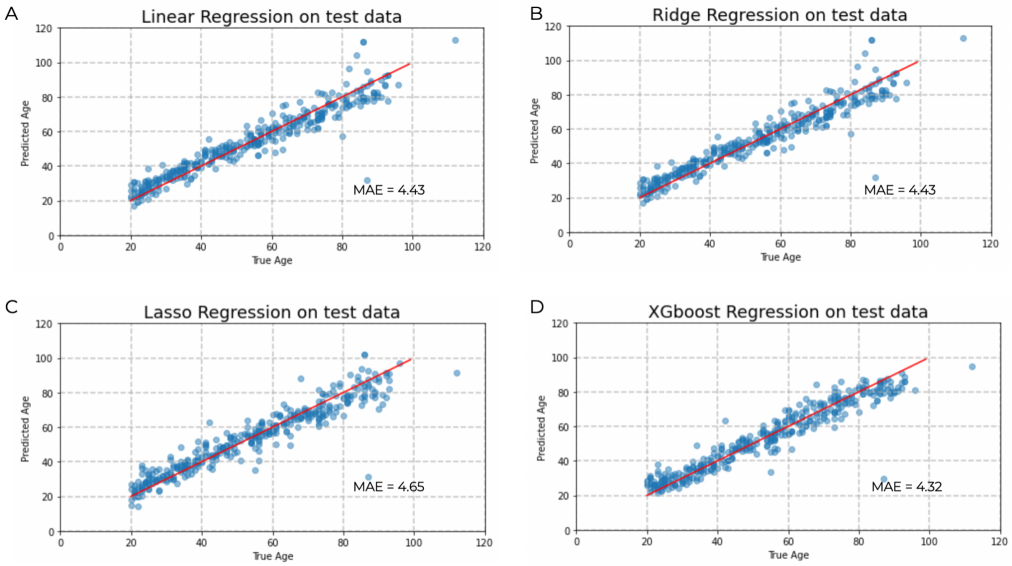


Figure 5. Predicting age from whole-blood DNA methylation data using all 375,603 features, 1066 samples.

4.2.1.1 Feature selection

The number of features (p) used in the models used above is far greater than the number of samples (n), a situation that promotes overfitting. To mitigate this potential problem and to try to identify which cpg sites are most important for age prediction, we attempted to reduce this number of features in several ways: statistical testing of linear fits, partial least squares regression, a bootstrapped correlation analysis, the ranking of Shapley scores, and by using feature importances from XGboost regression. Of these, the first and fourth and fifth worked

best and yielded similar results. We proceeded with the XGboost method. This worked as follows. We optimized an XGboost model's hyperparameters using all cpg sites and our training data. We then ran the model on randomly chosen 80/20 train-validate splits of the data. We did this 50 times and recorded which cpgs most often appeared in the models' top 100 most important features. On this basis, the cpg sites were then ranked in order of importance and modeling was done with the top-ranked sites.

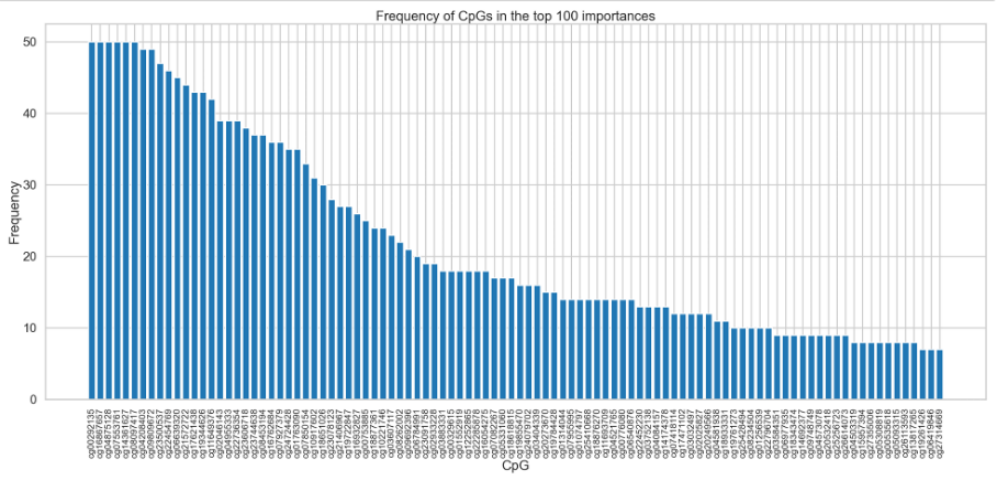


Figure 6. Frequency of cpgs occurring in the top 100 importance scores

Above is a histogram produced in this way (Fig. 6). It shows the cpgs that occurred most often among the top importance scores. Note, five cpgs appeared in all 50 trials, and many others also appeared repeatedly, a result that is extremely unlikely to occur by chance. Indeed, the probability of any cpg showing up by chance more than 4 times in the 50 trials is p = 7.66e-7. Thus, this method is selective and presumably it selects for those cpgs whose methylation is most associated with aging.

Next we repeated the age-predictive modeling but now using only the top 100 most important cpg sites (Fig. 7). Remarkably, after cutting the models' features from over 400,000 to 100, the smaller models performed comparably, and this was not the case if a random set of 100 cpgs was used. This indicates that our cpg-ranking method has some merit, and that many of the cpg sites in the dataset sites are likely not relevant to age prediction.

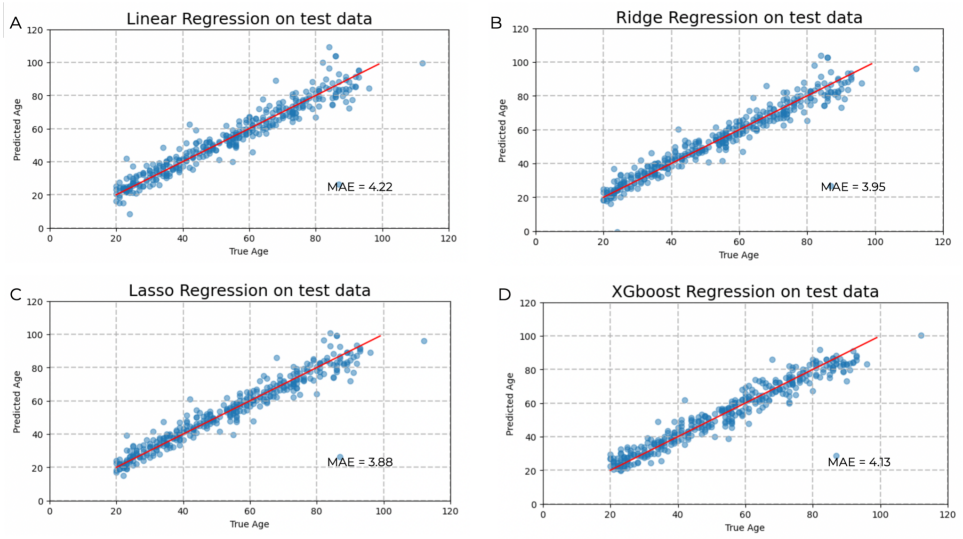


Figure 7. Predicting age using the top 100 cpg sites ranked by XGboost cross-validation.

Having had some success with 100 cpgs, we next investigated how many of the top-ranked cpgs is optimal for this type of modeling. To do this, for each model we repeatedly split the training data 80/20 using a fixed number of cpgs. We did this 50 times and fit the validation set each time. Then, this was repeated for different numbers of cpgs. The results are shown in Fig. 6 below. Most interesting, the optimal number of cpgs starts to plateau at ~100 for un-regularized Linear regression(Fig. 8-A) and ~1000 for the Ridge, Lasso, and XGboost models (Fig. 8 B-D).

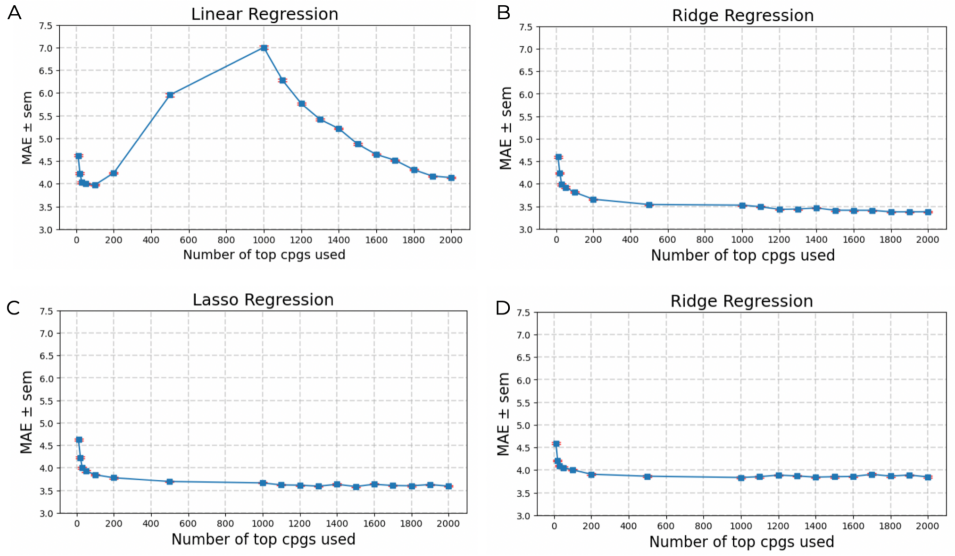


Figure 8. Mean absolute error as a function of the number of ranked cpgs used.

Thus, we repeated the modeling with now the top 1000 cpgs. (Fig. 9). Of note, the improvement over models built with top 100 cpgs (Fig. 7) was modest, however, the best model to this point was then a ridge regression model which used the top-ranked 1000 cpgs and had a MAE of 3.73 years..

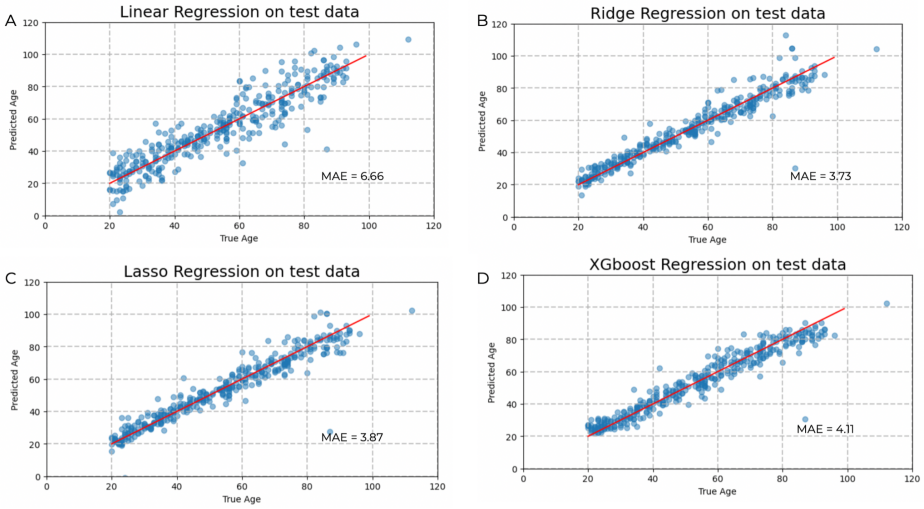


Figure 9. Predicting age using the top 1000 cpg sites ranked by XGboost cross-validation.

We then explored whether we could predict age more accurately with neural network models. We started by again testing how many features would be optimal. Twenty trials were performed with each of an increasing series of top-ranked cpgs, and model error on each validation set was recorded. The results are shown in Fig. 10. Specifically, we started with a preliminary neural network structure of 3 layers (hidden layer node number 128->56>28) and 2 layers (hidden layer node number 128->56).

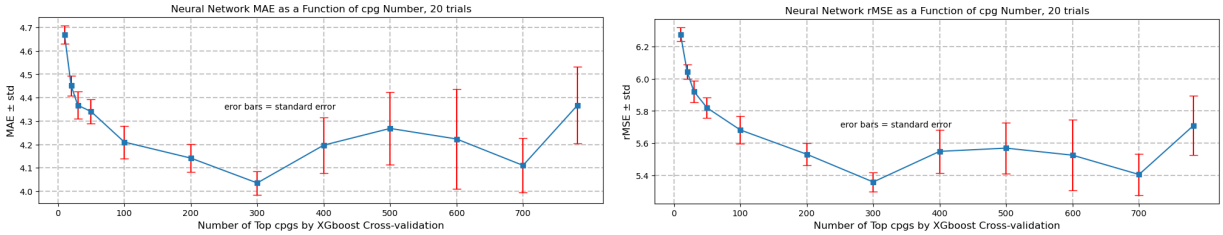


Figure 10-A. Mean MAE (left) & Mean rMSE (right) with different number of top features for Neural Network with 3 hidden layers over 20 trials

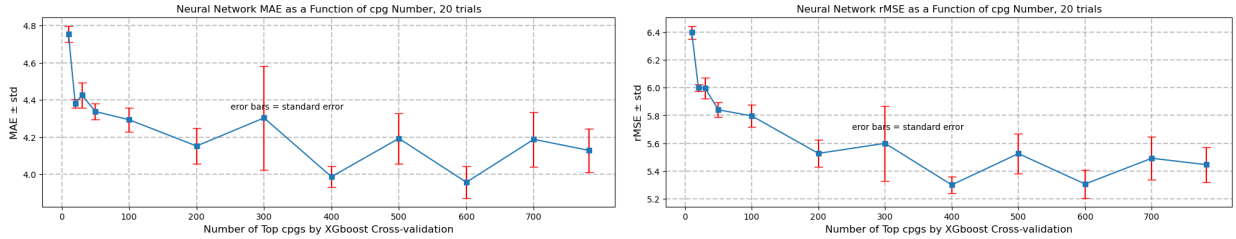


Figure 10-B. Mean MAE (left) & Mean rMSE (right) with different number of top features for Neural Network with 2 hidden layers over 20 trials

With the guidance from Fig. 10, we found that neural networks with 3 hidden layers perform best with between 300 and 700 cpgs, subject to fluctuations due to the randomness inherent in the models (for example, in the weight initializations etc) (Fig. 10-A). And the performance of models with 2 hidden layers generally plateaued at ~400 cpgs and above(Fig. 10-B). With this information, we continued to refine our neural network models by doing stepwise

refinement on hidden layer node number and node activation methods within such input feature range. After refinement, were able to achieve an MAE of 3.597 years and an rMSE of 4.841 years with 2 hidden layer neural networks (hidden layer node number 128->64) with input dimension of 700 cpgs (Fig. 11).

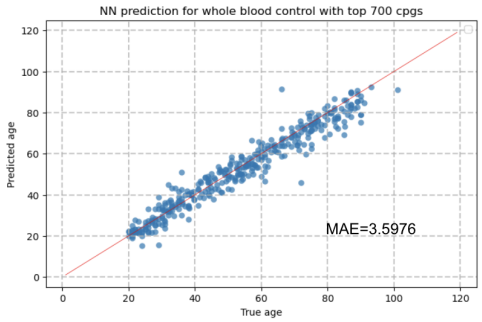


Figure 11. Predicting age using Neural Network with 3 hidden layers with 700 top cpgs

Table 12 summarizes our modeling results with DNA methylation data from whole blood from a healthy cohort. Our model with the smallest error is the neural network whose performance is shown above. It is composed of 2 hidden layers, 128 nodes in the first layer and 64 nodes in the second. It employs 700 of the top ranked sites ranked by XGBoost cross validation. Comparing these results to the literature. The error of our model is comparable to that of Horvath (2012) and Hannum (2012) and not as good as that of Zhang et al (2019), who reported a rMSE on some datasets as low as 2.04 years.

Model	MSE	rMSE	MAE	r^2	Corr
1000 cpgs					
Linear	88.680	9.417	6.669	0.803	0.912
Ridge	36.610	6.051	3.733	0.918	0.959
Lasso	36.830	6.609	3.866	0.918	0.958
XGboost	34.580	5.880	4.109	0.923	0.961
100 cpgs					
Linear	41.635	6.453	4.224	0.907	0.953
Ridge	37.580	6.130	3.950	0.916	0.957
Lasso	37.510	6.125	3.881	0.916	0.957
Xgboost	35.380	5.948	4.126	0.921	0.960
700 cpgs					
Neural Net	23.470	4.841	3.597	—	—

Table 12. Summary of models fitted to DNA methylation data from blood (test data).

4.2.2 Age Prediction Modeling - Are models transferable across different features?

We also explored whether we could take the models we built with DNA methylation data from blood and apply them without modification using methylation data from other tissues. As shown below in Fig. 13, however, this was not successful. When our blood-fitted ridge model is applied to methylation data from the brain, its age predictions are flat and close to 40 years of age. (Fig. 13-A). And when it is applied to data from breast tissue, its predictions are again flat but now close to 80 years of age (Fig. 13-B).

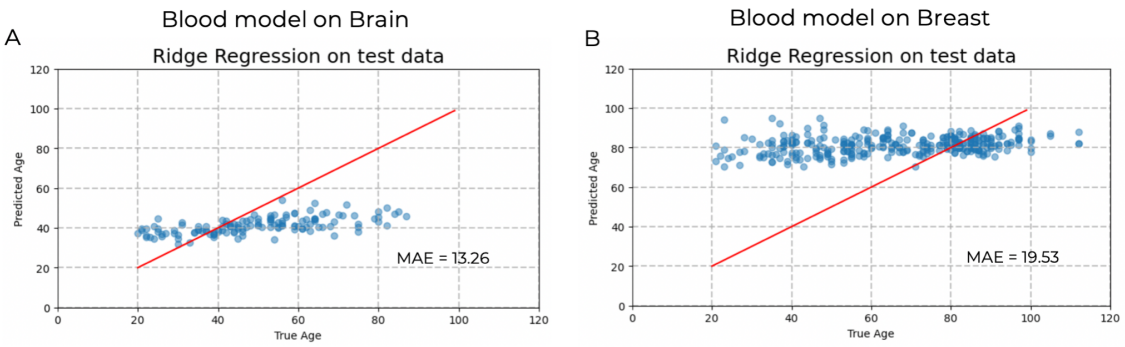


Figure 13. Apply ridge models developed with whole-blood DNA methylation data using data from other tissues (1000 cpgs).

Similar systematic prediction shift was also observed when predicting age using blood-trained neural network models (Fig. 14). We see a general underprediction when the blood-fitted neural network is applied to methylation data from brain tissue and an overprediction when it is applied to breast data although not as severe compared to that of the ridge regression.

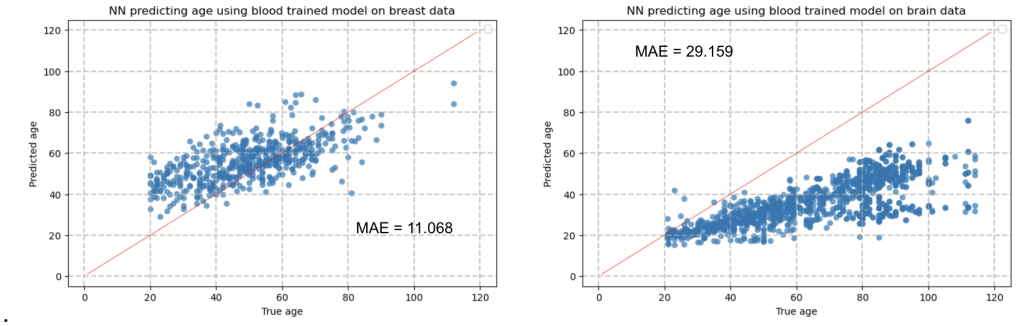


Figure 14. Apply Neural network models developed with whole-blood DNA methylation data using data from other tissues.

Thus, the models described above are not transferable without modification to these tissues, and it may be that cpg sites whose methylation correlates with age in one tissue does not in another.

4.2.3 Age Prediction Modeling - Are features transferable across different tissues?

While transferring the blood-based models without modification to other tissues did not work well, we investigated whether the top-ranked features generated XGBoost cross validation using methylation data from blood are valuable at all for predicting age with methylation data from other tissues, and as shown below, we did find that some of the features identified as important with the blood data could be used effectively in age-predictive modeling with leukocyte, breast and brain methylation data (transferability in features).

As above, we followed the procedure of first testing model performance on different numbers of cpgs as input features and then building models accordingly. The first tissue

we tested was Leukocytes (Fig. 15). Here 782 cpgs was found to be optimal, and we were able to generate a neural network model with the 782 top-ranked blood cpgs as input, and 2 hidden layers (hidden layer node number 128->56), that was able to achieve a rMSE of 4.412 and MAE of 3.510. On the brain samples. Remarkably, this performance is better than our best blood model.

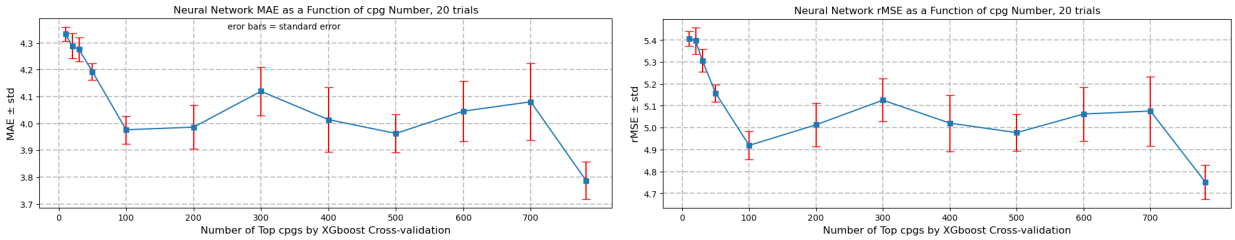


Figure 15-A. Mean MAE (left) & Mean rMSE (right) with different numbers of top whole blood features for Neural Network trained and tested with leukocyte methylation data.

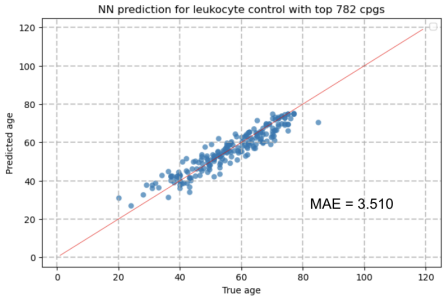


Figure 15-B. Predicting age using Neural Network fitted on leukocyte data with 2 hidden layers with 782 top cpgs generated from whole blood cross validation

Similar experiments were also conducted on breast data (Fig. 16), which yielded a neural network model with a rMSE of 7.897 and a MAE of 5.967; and on brain data (Fig. 17), which yielded a model with a rMSE of 7.962 and a MAE of 6.015 as shown below.

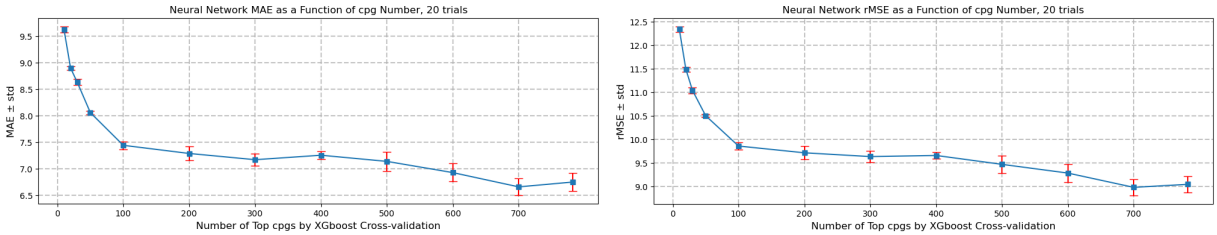


Figure 16-A. Mean MAE (left) & Mean rMSE (right) with different numbers of top whole blood features for Neural Network trained and tested with brain methylation data.

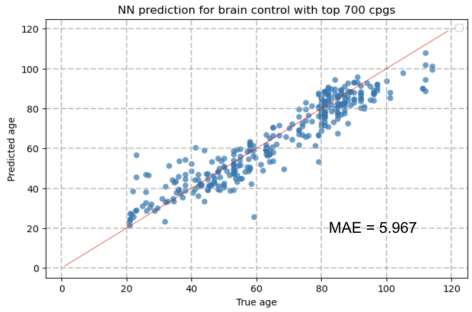


Figure 16-B. Predicting age using Neural Network fitted on brain data with 2 hidden layers with 500 top cpgs generated from whole blood cross validation

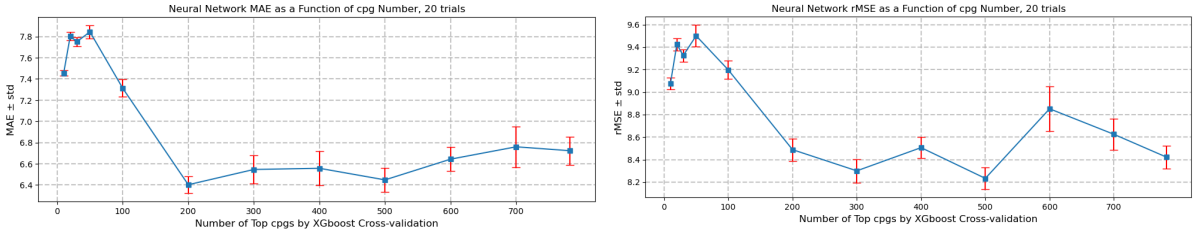


Figure 17-A. Mean MAE (left) & Mean rMSE (right) with different numbers of top whole blood features for Neural Network trained and tested with breast methylation data.

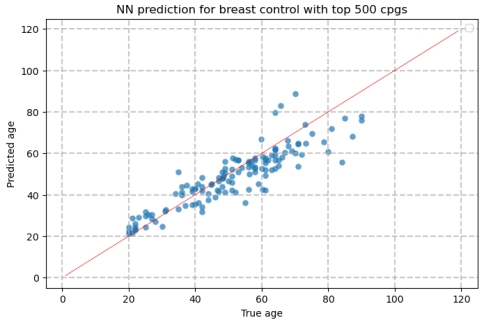


Figure 17-B. Predicting age using Neural Network fitted on breast data with 3 hidden layers with 500 top cpgs generated from whole blood cross validation

Although the model performance in terms of MAE and rMSE for brain and breast weren't comparable to that of leukocytes, they still yielded reasonable performance as shown in Fig. 16-B and Fig. 17-B. This means that certain cpgs, out of the over 400,000 in the blood dataset, can be used as general predictors across many tissues. Specifically, using the top CPGs generated by whole blood transfers the best on Leukocyte methylation. Such findings would significantly save us time in feature selection when studying different tissues.

4.2.4 Age Prediction Modeling - Brain, Breast Tissue:

In addition to simply co-opting the top-ranked cpgs identified with blood data for use in other tissues, we also specifically selected cpgs —using our XGboost cross validation technique— for each tissue separately. Models built with these features are shown for DNA methylation data derived from brain and breast tissue in Fig. (18). Interestingly, models built with data from these tissues are not as good at predicting age as are models built from DNA

methylation data from blood.(Compare FIg. 18-A, 18-B to 18-C). That is, their MAEs and rMSEs are substantially larger.

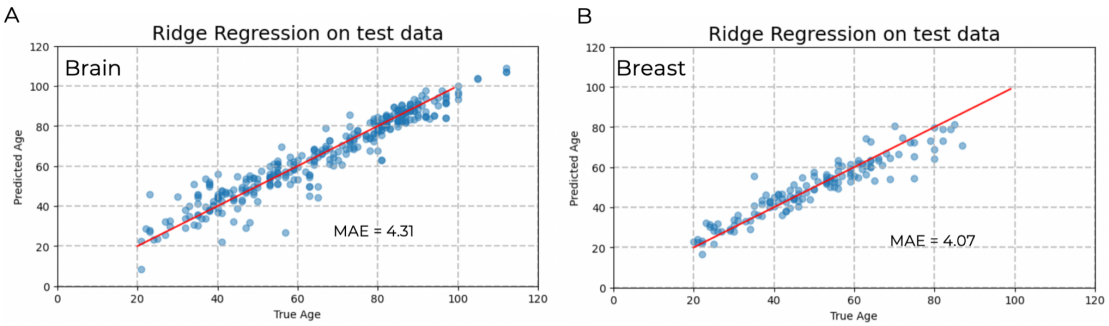


Figure 18. Linear models developed with the top 1000 cpgs from brain and breast data

With neural network models, we also did age prediction with brain and breast and found that for both, increasing numbers of top cpg features would improve test performance. A procedure similar to that of 4.2.1.1 is employed where we use the graphs in Fig. 19 and Fig. 20 as general guidance for a rough optimal input feature number. Then we do refinement on model structure to obtain the best performing model.

For brain neural network, 680 input cpgs had an rMSE of 6.33 and MAE of 4.479 which was similar to that of ridge regression.

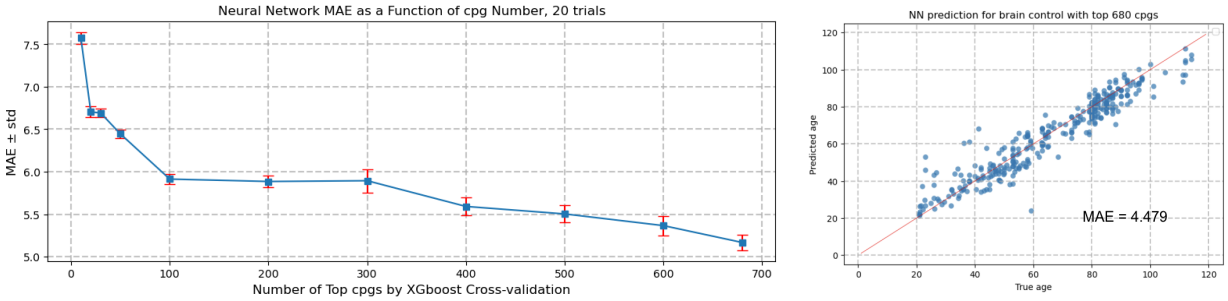


Figure 19. Neural Network models for all brain tissues developed with top 680 cpgs from brain XGboost CV rankings

For breast neural network, a model with 680 input cpgs had an rMSE is 7.38 and MAE of 5.56, which was slightly worse than that of ridge regression.

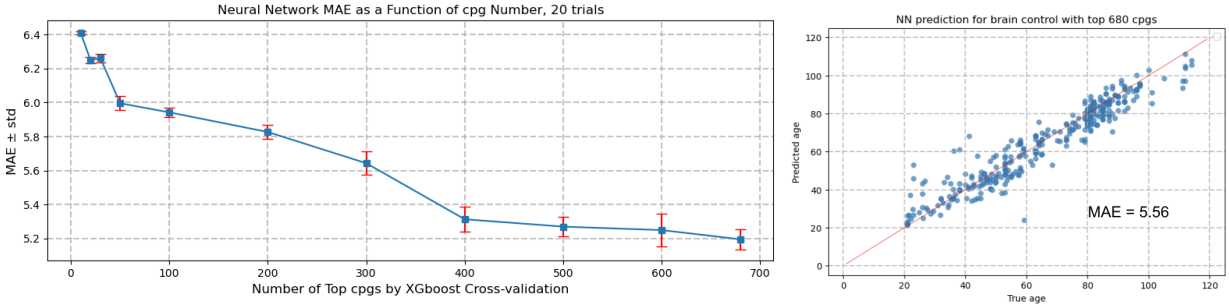


Figure 20. Neural Network models for all breast tissues developed with top 680 cpgs from breast XGboost CV rankings

4.3 Examining healthy vs unhealthy Cohort

This section focuses on determining whether the healthy age prediction model is applicable to unhealthy individuals, where unhealthy individuals are again defined as those with neurodegenerative disorders such as Alzheimer's, Parkinson's and Huntington's. The analysis was conducted on individuals from the unhealthy cohorts detailed in Figure 1 with the corresponding age distributions shown in Figure 3. Due to the small sample sizes for the blood tissue cohorts (for both Alzheimer's and Parkinson's) this section only presents the results for the brain tissue cohorts.

4.3.1 Transferring healthy models to unhealthy cohorts: brain model

The first point of analysis was to determine whether the tissue specific models developed in the previous sections are directly transferable to unhealthy individuals. In this section we present the analysis for brain tissue data on Alzheimer's and Huntington's patients.

Looking at the brain tissue data, 55 of the 100 top CpG sites identified during the healthy control cohort brain tissue feature selection were present in the unhealthy cohorts dataset. Consequently, the brain tissue models were retrained on the healthy control data using only these 55 CpG sites. The results can be seen in Figure 21. The performance of the model with 55 CpG sites was significantly lower than for 100 CpGs with MAE for ridge regression increasing from 4.31 to 5.54.

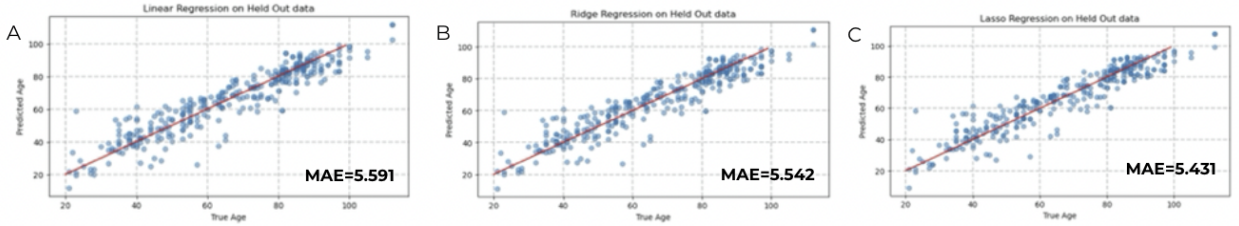


Figure 21. Performance of the brain tissue regression models trained on the healthy control subjects using 55 of the 100 top CpG sites. The results shown are for the held out data.

Applying these models directly to the Alzheimer's and Huntington's subjects we have the results shown in Figure 22. Across both the Alzheimer's and Huntington's cohorts we can see that the healthy model systematically overpredicts the ages. Due to this we can see much larger MAE values for both of the unhealthy groups for all 3 models. For instance an increase from MAE=5.591 for the healthy control cohort to MAE=51.371 for the Alzheimer's cohort for the Linear Regression model. Although the resulting statistics indicate that the healthy control brain models aren't directly applicable to the unhealthy cohorts, the strong correlation between the over predicted values indicates that the CpG sites themselves may be transferable but that the model weights need to be retrained.

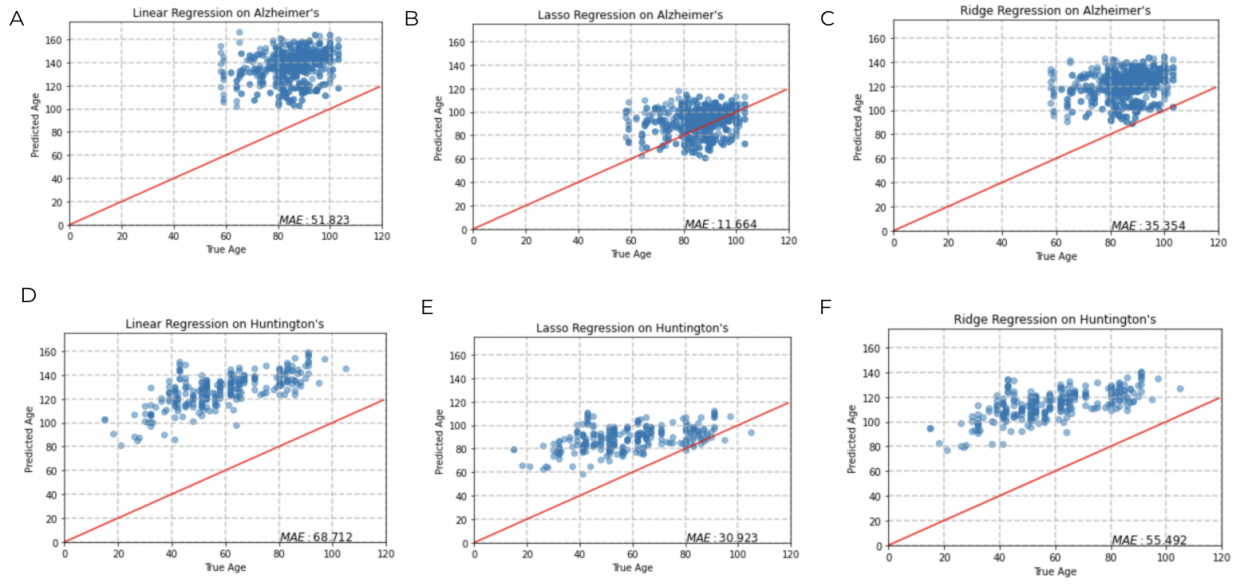


Figure 22. . Brain tissue healthy control models for 55 CpG sites applied to the A)-C) Alzheimer's brain tissue data (811) and D)-F) Huntington's brain tissue data (270).

4.3.2 Transferring significant healthy CpG sites to unhealthy cohorts

To further investigate whether the CpG sites are transferable between healthy and unhealthy cohorts, we retrain the regression models using the 55 CpG sites on the unhealthy data using the same methodology as outlined above.

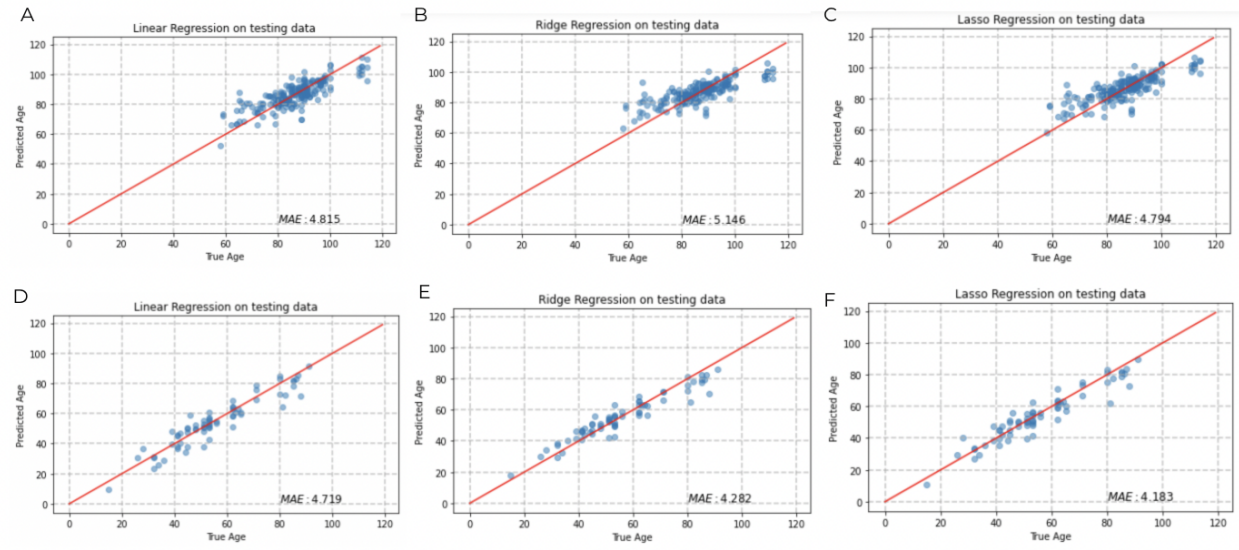


Figure 23. . Results of the brain tissue model on unhealthy cohorts A)-C) Alzheimer's D)-F) Huntington's using the 55 significant CpG sites from healthy cohorts applied to the test set

From Figure 23 we can see that retraining the model and obtaining new weights drastically improves the accuracy of the predictions. Interestingly, these perform better than the healthy model on healthy individuals despite the 55 features being based on the healthy cohort data. For the linear regression models, the MAE was 5.591 for the healthy model, 4.815 for the Alzheimer's cohort and 4.719 for Huntington's. For ridge regression we can see that the Huntington's cohort model achieved a MAE of 4.282, which is slightly better than the full healthy brain model which achieved a MAE of 4.31. Across the three models, the MAE for Huntington's was lower than for Alzheimer's patients. This difference could be due to underlying biological factors relating to the selected CpG sites, however one possible alternative explanation is the distribution of ages in these cohorts. From the distributions in Figure 3 C) and D) we can see that the Alzheimer's ages are skewed towards older individuals with the ages for Huntington's being fairly evenly distributed with a good range. It has been noted in literature that age predictive models using methylation data perform worse for much older individuals due to the CpG sites becoming saturated at those ages. Consequently, the majority of the individuals in the Alzheimer's cohort being much older could be the reason for slightly poorer model performance.

These results indicate that the features from the same tissue are transferable across healthy and unhealthy cohorts despite the models not being transferable. The nature of linear models also allows for a great degree of interpretability. Figure 24 shows a plot of the weights associated with each CpG site for the linear regression models for healthy control, Alzheimer's and Huntingston's cohorts. Different CpG sites have been associated with various gene productions, which is discussed in more detail in section 4.6. From Figure 24 we can see that for many of the CpG sites, not only the magnitude but also the sign of the weights associated with a CpG site change. For instance for cg04098194 the healthy control model had a weight of around 10, Huntington's model of around -50 and Alzheimer's close to -100.

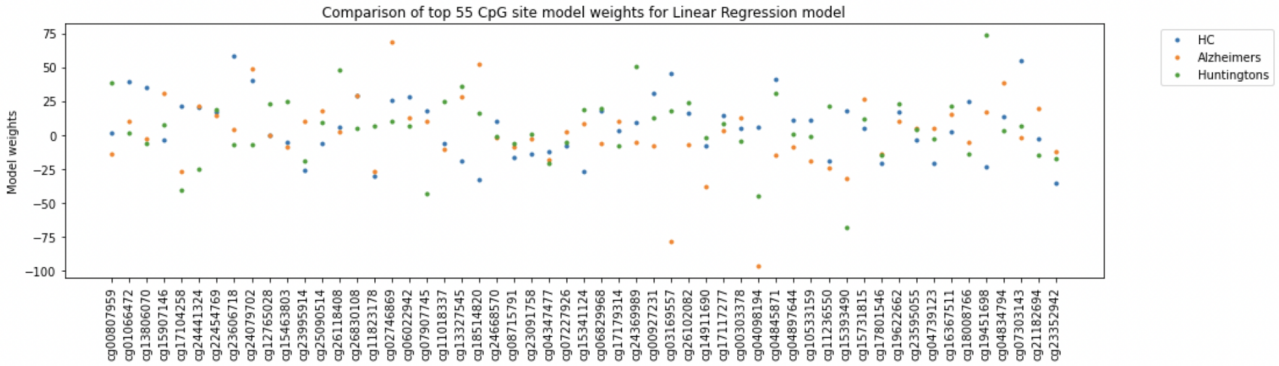


Figure 24. Plot of the linear regression model's weights for each of the 55 CpG sites for each of the healthy control, Alzheimer's and Huntington's models.

4.3.3 Classification model for healthy vs unhealthy

These significant differences in feature weights between healthy and unhealthy cohorts (whilst achieving comparable model performance across the cohorts) leads to the question of

whether it's possible to distinguish between healthy and unhealthy individuals based on the data from these 55 CpG sites. We fit a logistic regression model to the brain tissue data with the healthy control individuals being classes as healthy and Alzheimer's and Huntington's cohorts being pulled into one unhealthy class. Figure 25 shows the results for the test set. We can see that the model achieved a class accuracy of 0.68, meaning that 68% of the individuals are correctly assigned to their respective healthy and unhealthy classes.

	precision	recall	f1-score	support
Unhealthy	0.66	0.63	0.65	271
Healthy	0.70	0.72	0.71	322
accuracy			0.68	593
macro avg	0.68	0.68	0.68	593
weighted avg	0.68	0.68	0.68	593

Figure 25. Results for the test set of the logistic regression classification model of healthy and unhealthy subjects using the 55 CpG sites as features.

4.4 Biological significance

With regard to our results, a natural question is which genes might methylation be affecting and thereby perhaps affecting aging? Table 26 below maps the top-ranked 100 cpg sites in blood to their closest genes, and shown in Fig. 27 is a histogram that demonstrates that some genes are associated with more than one of the top-ranked cpgs. Five of the top 100 cpgs, for example, are associated with the KLF14 gene, a transcription factor thought to be a master regulator of gene expression in adipose tissue.

cpg	gene	cpg	gene	cpg	gene	cpg	gene
cg14361627	KLF14	cg07927379	C7orf13	cg18933331		cg04084157	VGF
cg16867657	ELOVL2	cg19722847	IPO8	cg17471102	FUT3	cg10149533	
cg24724428	ELOVL2	cg10917602	HSD3B7	cg20010135	HSD3B7	cg17110586	
cg11649376	ACSS3	cg21406967	TRIP6	cg25256723	F5	cg09499629	KLF14
cg24079702	FHL2	cg09692396	LRRC23	cg06540876	ZBTB12	cg25428494	HPSE
cg04875128	OTUD7A	cg16762684	MBP	cg12580096	C19orf57	cg09748749	ASL
cg08097417	KLF14	cg01763090	OTUD7A	cg11693709	PAK6	cg04503319	ANKRD11
cg00292135	C7orf13	cg23078123	GPR177	cg19784428	NWD1	cg20249566	NWD1
cg02046143	IGSF9B	cg25410668	RPA2	cg01256539	PRR16	cg25693132	GRM2
cg07553761	TRIM59	cg07082267		cg04521765	LOXL4	cg11220950	SYNGR3
cg21572722	ELOVL2	cg02933228	CDC42BPG	cg01314044		cg00808969	USP35
cg04208403	ZNF423	cg23606718	FAM123C	cg22285878	KLF14	cg03752138	SOCS3
cg23500537		cg07955995	KLF14	cg01074797	PDZK1IP1	cg09648727	
cg08262002	LDB2	cg05331060		cg10943497	MEG3	cg15957394	AFAP1
cg04955333	IQCE	cg18651026	COL11A2	cg03032497		cg16008966	
cg09809672	EDARADD	cg10221746		cg19855470	CACNA1I	cg21186299	VGF
cg06639320	FHL2	cg05308819		cg00776080	TENC1	cg20273670	
cg17621438	RNF180	cg18618815	COL1A1	cg16054275	F5	cg18725681	FITM2
cg22736354	NHLRC1	cg18877361		cg23091758	NRIP3	cg22016779	DNER
cg22454769	FHL2	cg12252865	HDAC11	cg01552919	GAK	cg01676322	ACBD4
cg19344626	NWD1	cg16932827		cg04581938		cg21296230	GREM1
cg23744638		cg03883331		cg03404339	KRT7	cg26614073	SCAP
cg07850154	RNF180	cg07135942	ZNF238	cg00003345	CASZ1	cg01014399	
cg08453194	CCND3	cg03607117	SFMBT1	cg02025827	HSD3B7	cg06784991	ZYG11A
cg07927379	C7orf13	cg00753885		cg22796704	ARHGAP22	cg18343474	

Table 26. Genes associated with the top 100 ranked cpg sites from blood

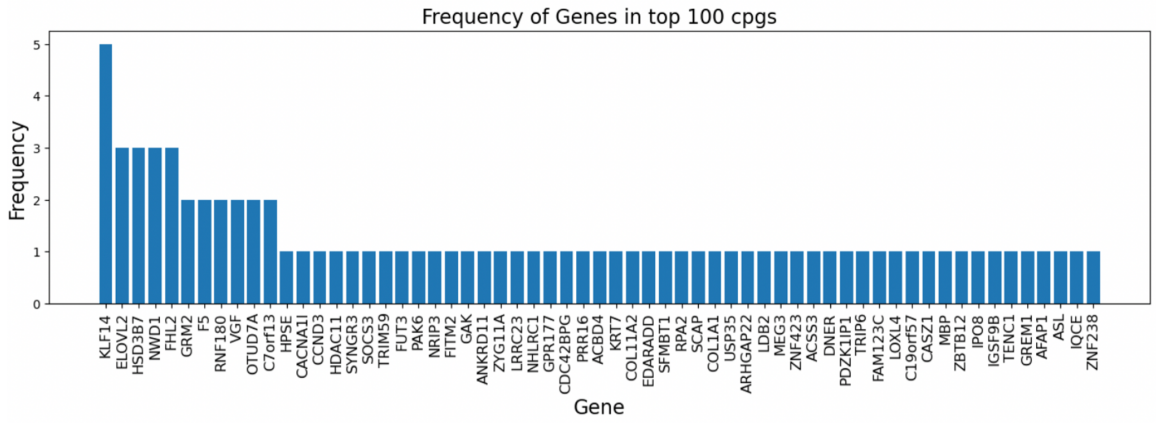


Fig 27. Genes associated with the top 100 ranked cpg sites from blood

Table 28 below focuses on the top 20 highest ranked cpg sites and their associated genes. Four properties stand out. First, like KLF14, two other genes are also associated with fat cells or fat metabolism, ELOVL2 and ZNF423 (blue). Thus, it may be that processes involving fat metabolism and storage have an important influence on aging. Second, there are four genes associated with the ubiquitin-proteasome pathway (red), OTUD7A, TRIM59, RNF180, and NHLRC1, an important pathway for protein degradation. In fact three of these genes are E3 ubiquitin ligases, which are responsible for marking proteins for degradation. Thus, in terms of looking for interventions in the aging process, targeting this pathway may be a promising avenue of investigation. Indeed, irrespective of DNA methylation, several studies had identified this pathway as having an important influence on aging (3, 13). Third, many of the genes in Table 28 contain a structure known as a zinc finger, a structure that binds Zn^{2+} and is often involved in DNA-protein and protein-protein interactions. The significance of this observation is as yet unclear. And, finally, nearly all of the methylation sites in the table have been identified in other studies as being related to aging (right-most column). Thus, our results are in accord with the growing literature on DNA methylation and aging.

Rank	cpg	Gene	Function	Zinc finger	Refs related to aging
1	cg14361627	KLF14	Krüppel-Like Factor 14 (KLF14), transcription factor, master regulator of gene expression in the adipose tissue	x	16, 8, 5, 7
2	cg16867657	ELOVL2	Fatty Acid Elongase 2, involved in the synthesis of very long polyunsaturated fatty acids		21, 15, 14, 17,5, 7
3	cg24724428	ELOVL2			15, 14, 17,5, 7
4	cg11649376	ACSS3	Acyl-CoA Synthetase Short Chain Family Member 3, Ligates acetate and CoA6		1
5	cg24079702	FHL2	Four And A Half LIM Domains 2, Assembly of extracellular membranes, double zinc finger, LIM protein	x	5,17,2,
6	cg04875128	OTUD7A	OTU Deubiquitinase 7A, deubiquitinizing enzyme and possible tumor suppressor, zinc finger	x	21, 17,7
7	cg08097417	KLF14		x	21, 16, 8, 5, 7
8	cg00292135	C7orf13	Not much known		
9	cg02046143	IGSF9B	Immunoglobulin Superfamily Member 9B, cell adhesion, localized to inhibitory synapses		21, 7
10	cg07553761	TRIM59	Tripartite Motif Containing 59, E3 ubiquitin ligase, zinc finger, RING finger protein	x	15, 7
11	cg21572722	ELOVL2			15, 14, 17,5, 7
12	cg04208403	ZNF423	Zinc Finger Protein 423, Krüppel-Like Factor, zinc finger transcription factor, KO affects adipogenesis	x	16
13	cg23500537				21
14	cg08262002	LDB2	LIM Domain Binding 2, adapter molecule, binds LIM		14,15
15	cg04955333	IQCE	IQ Motif Containing E, signaling by GPCR and Hedgehog		21
16	cg09809672	EDARADD	EDAR Associated Death Domain, Ectodysplasin-A receptor-associated adapter protein		21, 16, 4, 9
17	cg06639320	FHL2		x	21, 5,17,2,
18	cg17621438	RNF180	E3 Ubiquitin-Protein Ligase RNF180, promotes protein degradation by the proteasome pathway	x	21
19	cg22736354	NHLRC1	E3 Ubiquitin-Protein Ligase NHLRC1 , promotes protein degradation by the proteasome pathway	x	9
20	cg22454769	FHL2		x	21, 5,17,2,
21	cg19344626	NWD1	NACHT And WD Repeat Domain Containing 1, modulator of androgen receptor activity		
22	cg23744638				21
23	cg07850154	RNF180		x	

Table 28. Genes associated with the top 23 ranked cpg sites, blood data.

Finally, it may be of interest to know which of the cpg sites that we have identified as useful for chronological age prediction are found in other such models. In the literature there are principally three such models: Horvath(2013), Hannum et al. (2013), and Zhang et al.(2019). Table 29 below shows the number of cpgs these models have in common with each other and with our top 100 sites. Most informative is the last column Itindicates that of Horvath's 353 model cpg sites, Hannum's 71 sites, and Zhang's 514 sites, 6, 28 and 37 are found in our top 100 respectively.

	Horvath	Hanuum	Zhang	us top 100
Horvath	353	6	11	6
Hannum	6	71	30	28
Zhang	11	30	514	37
us top 100	6	28	37	100

Table 29. Our cpg sites common to the Horvath, Hannum and Zhang models

Text References

Salameh Y, Bejaoui Y and El Hajj N (2020) DNA Methylation Biomarkers in Aging and Age-Related Diseases. Front. Genet. 11:171.

Weidner, C., Lin, Q., Koch, C., Eisele, L., Beier, F., Ziegler, P., et al. (2014). Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol. 15:R24.

Horvath, S. (2013). DNA methylation age of human tissues and cell types. Genome Biol. 14:R115. doi: 10.1186/gb-2013-14-10-r115

Hannum, G., Guinney, J., Zhao, L., Zhang, L., Hughes, G., Sadda, S., et al. (2013). Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol. Cell 49, 359–367. doi: 10.1016/j.molcel.2012.10.016

Langner, Taro, Johan Wikström, Tomas Bjerner, Håkan Ahlström, and Joel Kullberg. "Identifying morphological indicators of aging with neural networks on large-scale whole-body MRI." IEEE transactions on medical imaging 39, no. 5 (2019): 1430-1437.

Tozer DJ, Zeestraten E, Lawrence AJ, Barrick TR, Markus HS. Texture Analysis of T1-Weighted and Fluid-Attenuated Inversion Recovery Images Detects Abnormalities That Correlate With Cognitive Decline in Small Vessel Disease. Stroke. 2018 Jul;49(7):1656-1661.

Putin E, Mamoshina P, Aliper A, Korzinkin M, Moskalev A, Kolosov A, Ostrovskiy A, Cantor C, Vijg J, Zhavoronkov A. Deep biomarkers of human aging: Application of deep neural networks to biomarker development. Aging (Albany NY). 2016 May;8(5):1021-33.

Frenk, S. and J. Houseley, Gene expression hallmarks of cellular ageing. Biogerontology, 2018. 19(6): p. 547-566.

Zhang, Q., et al., Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. Genome Med, 2019. 11(1): p. 54.

Table References

1. Arpon, A., et al., Methylome-Wide Association Study in Peripheral White Blood Cells Focusing on Central Obesity and Inflammation. Genes (Basel), 2019. 10(6).
2. Bacalini, M.G., et al., Systemic Age-Associated DNA Hypermethylation of ELOVL2 Gene: In Vivo and In Vitro Evidences of a Cell Replication Process. J Gerontol A Biol Sci Med Sci, 2017. 72(8): p. 1015-1023.
3. Bergsma, T. and E. Rogaeva, DNA Methylation Clocks and Their Predictive Capacity for Aging Phenotypes and Healthspan. Neurosci Insights, 2020. 15: p. 2633105520942221.
4. Bocklandt, S., et al., Epigenetic predictor of age. PLoS One, 2011. 6(6): p. E14821.
5. Bysani, M., et al., Epigenetic alterations in blood mirror age-associated DNA methylation and gene expression changes in human liver. Epigenomics, 2017. 9(2): p. 105-122.
6. Florath, I., et al., Cross-sectional and longitudinal changes in DNA methylation with age: an epigenome-wide analysis revealing over 60 novel age-associated CpG sites. Hum Mol Genet, 2014. 23(5): p. 1186-201.
7. Hannum, G., et al., Genome-wide methylation profiles reveal quantitative views of human aging rates. Mol Cell, 2013. 49(2): p. 359-367.
8. Hegde, A.N., et al., Perturbations of Ubiquitin-Proteasome-Mediated Proteolysis in Aging and Alzheimer's Disease. Front Aging Neurosci, 2019. 11: p. 324.
9. Hong, S.R., et al., DNA methylation-based age prediction from saliva: High age predictability by combination of 7 CpG markers. Forensic Sci Int Genet, 2017. 29: p. 118-125.

10. Horvath, S., DNA methylation age of human tissues and cell types. Genome Biol, 2013. 14(10): p. R115.

11. Huang, Y., et al., Developing a DNA methylation assay for human age prediction in blood and bloodstain. Forensic Sci Int Genet, 2015. 17: p. 129-136.

12. Johansson, A., S. Enroth, and U. Gyllensten, Continuous Aging of the Human DNA Methylome Throughout the Human Lifespan. PLoS One, 2013. 8(6): p. E67378.

13. Kevei, E. and T. Hoppe, Ubiquitin sets the timer: impacts on aging and longevity. Nat Struct Mol Biol, 2014. 21(4): p. 290-2.

14. Koch, C.M. and W. Wagner, Epigenetic-aging-signature to determine age in different tissues. Aging (Albany NY), 2011. 3(10): p. 1018-27.

15. Levy, J.J., et al., MethylNet: an automated and modular deep learning approach for DNA methylation analysis. BMC Bioinformatics, 2020. 21(1): p. 108.

16. Marttila, S., et al., Ageing-associated changes in the human DNA methylome: genomic locations and effects on gene expression. BMC Genomics, 2015. 16: p. 179.

17. Naue, J., et al., Chronological age prediction based on DNA methylation: Massive parallel sequencing and random forest regression. Forensic Sci Int Genet, 2017. 31: p. 19-28.

18. Pan, C., et al., The evaluation of seven age-related CpGs for forensic purpose in blood from Chinese Han population. Forensic Sci Int Genet, 2020. 46: p. 102251.

19. Tharakan, R., et al., Blood DNA Methylation and Aging: A Cross-Sectional Analysis and Longitudinal Validation in the InCHIANTI Study. J Gerontol A Biol Sci Med Sci, 2020. 75(11): p. 2051-2055.

20. Weidner, C.I., et al., Aging of blood can be tracked by DNA methylation changes at just three CpG sites. Genome Biol, 2014. 15(2): p. R24.

21. Zhang, M., et al., DNA methylation age acceleration is associated with ALS age of onset and survival. Acta Neuropathol, 2020. 139(5): p. 943-946.

22. Zhang, Q., et al., Improved precision of epigenetic clock estimates across tissues and its implication for biological ageing. Genome Med, 2019. 11(1): p. 54.

Appendix:

Appendix A. Recap of Milestone 1:

For Milestone 1 a significant amount of work was devoted to exploration and identification of appropriate datasets and literature reviews in three areas: MRI images, blood chemistry, and DNA methylation data. After further discussion with Merck, the MRI image avenue was not taken further, as the MRI images available to the team were minimally processed with a large amount of computational effort being required to do so without a clear return. The initial exploration of these data types in the PPMI database, led to the conclusion that DNA methylation data was the most promising age predictor. Thus, working with this type of data has become our main focus (for more details, please refer to the Milestone 1 technique report). The following sections, therefore, describe what has been completed since Milestone 1 with a focus on the analyses we have done with DNA methylation data obtained from the EWAS Datahub. Refer to Fig. 1 below for a roadmap of the whole project and where we are now.

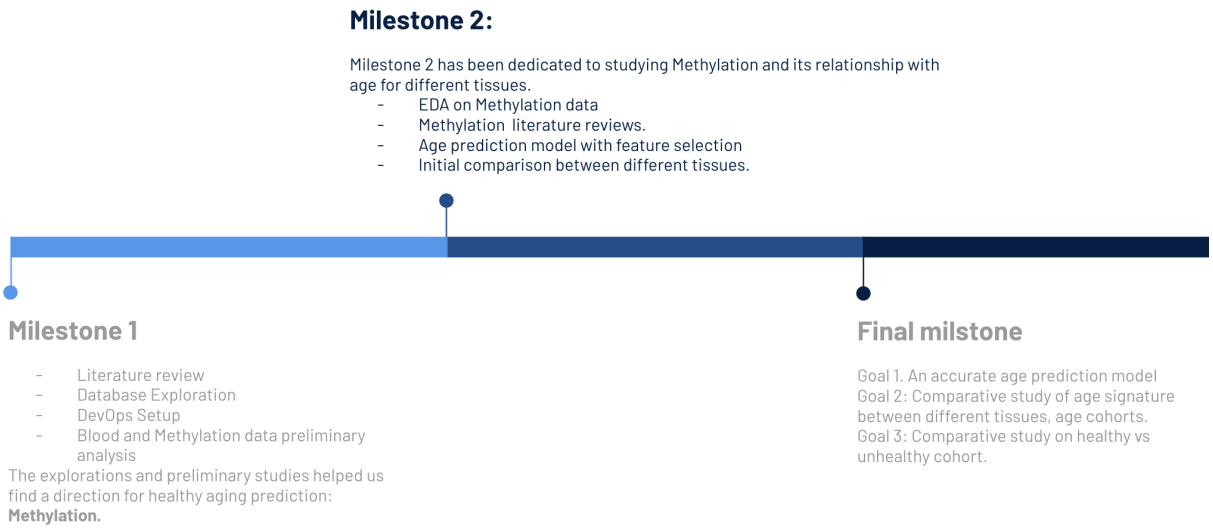


Figure 1. Road map of different milestones

Healthy Aging Signal Research
Summary Report Milestone 3

Partner:
Merck & Co.

Group Members:
Eleonora Shantsila: eshantsila@g.harvard.edu,
Yaxin Lei: yaxin_lei@g.harvard.edu,
Aaron Jacobson: aaronjacobson@g.harvard.edu ,
Daniel Cox: daniel_cox@g.harvard.edu

Hello Antong and Greg,

I hope you are well. Remarkably, it is time again for us to submit a technical report, this time for the last and final milestone of our project Milestone 3. The report is attached. It contains a brief new introduction and then a series of sections that describe much of the work we have done for this project. Perhaps of most interest, figures for all the previous Milestone 2 sections have been revised and in some places expanded to reflect the latest versions of our models. New work done to optimize the number of methylation sites used in our models has been added, and a good deal of new work has also been added, sections 4.1 through 4.5, that examines differences in DNA methylation between healthy and unhealthy cohorts. In addition, a section has been added that briefly discusses the DNA-methylation-and-aging literature, and a fairly extensive section, 4.6, has been added that discusses the genes associated with the methylation sites we have found to be most important for age prediction. Currently, we are working to convert our extensive code to a useful and deliverable form and to prepare the presentation, poster, and blog-post that we must also submit shortly.. We are looking forward to discussing with you this Thursday, our most recent results, how we might structure our final github repo, and what else we might do that would be useful to Merck. We thank you very much for your guidance over the past many weeks. We have enjoyed working on this project and learned a great deal. We greatly appreciate the time and energy you have given to our project and Merck's willingness to engage with students at Harvard.

Best Regards, ,

Daniel Cox
Eleonora Shantsilla
Yaxin Lei
Aaron Jacobson