Healthy Aging Signal Research
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

Problem Description

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

Specifically our goals are the following:

1. To Identify databases that we might use to identify biomarkers of healthy aging.
2. To identify biomarkers that are significantly related to age.
3. To build models that can predict healthy aging based on the biomarkers we identify.

We will consider this project successful if we are able to develop models that can predict age based on biomarkers to a rms accuracy of 5 years. Such a model will serve as a baseline against which the aging process in healthy individuals can be assessed and compared against the unhealthy population, and once armed with this knowledge, the need for therapeutic intervention to prevent the onset of disease can then be assessed.

Initial Preparation:

Our goal is to help our partner find important healthy aging signals through a wide range of datasets with various biological features. In order to do this, we first find as much relevant data as possible, then narrow down our selection of which to use for our predictive models. We will then build models to predict age from the chosen biomarkers, and compare them to models on unhealthy patients. As we are trying to predict biological age, we will train our primary models on the chronological age of healthy patients. In doing this, we hope to find specific indicators that denote aging. We plan to start with interpretable simpler models, then refine signal extraction through more complicated models with potential loss of interpretability.

- **Data:**

We were not given a specific dataset to work with, however, over the last few weeks we have been examining four databases Merck suggested might be of interest and looking for others.

- The first is the HCP-Aging human connectome database, which contains MRI brain images from 689 healthy subjects ranging in age from 36 to 100. This database might be useful, but preliminarily we have shifted our focus from it, as it was agreed that the amount of effort and computing power it would take to work with these images may not be worth the risk of small payoff.

- We have also considered two databases related to Alzheimer's disease (The ADNI Neuroimaging Initiative and The AD knowledge portal), but we have not examined these thoroughly because the age ranges of the control subjects in these studies are not expected to be wide enough to be of use to us.

- However, we have also investigated the PPMI Parkinson's disease database. This database does have individuals with a sufficiently wide age range for our purposes (30 to 100), and of the 600 total subjects in the database 200 are healthy controls. Also, this database contains a wide variety of potential age-related biomarkers including measurements of the levels of various blood and CSF molecules, brain scans, cognitive tests, vital signs, and considerable genetic data. At present, therefore, we consider this database the most promising for our work, and as discussed below, we have explored in some depth two types of PPMI biomarkers that the aging literature suggests might reasonably be related to age, blood chemistry and DNA methylation. Of these DNA methylation now looks most promising,

- Given the promising DNA-methylation EDA, most recently we have been looking for a database with a larger DNA methylation dataset than contained in the PPMI data, and we have found one that appears to be very good. It is called the Epigenome Wide Association Studies (EWAS) datahub. A preliminary search of this database indicates it has ~9000 DNA methylation samples that may meet our needs. We are exploring these samples now.

- **DevOps:**
  - To successfully conduct our collaborative work and document our processes, our initial preparation also includes familiarity and setup of the following DevOps items.
    - Github (main documentation and public use)
    - Google Drive (internal document collaboration)
    - AWS (internal computation use)
    - Deep Notes (facilitates initial EDA)

Literature Review

Our literature review identified three key areas of interest relating to predicting human age: using blood chemistry markers for the prediction of human chronological age (Putin et al 2015); using MRI images to extract patient features for age prediction (Tozer et al 2018 and Lagner et al 2020) and using DNA methylation age of various tissues (Horvath 2013). Below we outline the key features of these models.

Blood Chemistry

In their 2015 paper Application of deep neural networks to biomarker development Putin et al. developed a series of 21 DNNs (Deep Neural Networks) of varying depth, structure and optimisation methods to predict human chronological age from blood test results. The data from 62,419 healthy individuals was used with the records containing age, sex and results from 46 standard blood markers. Human age prediction was treated as a regression problem and 40 different DNN were trained, with the training set containing 56,177 individuals. Out of these the 21 best performing models were selected and stacked using standard coefficient of determination R^2 and epsilon prediction as two methods to estimate performance. The best single DNN performed with 0.8 R^2 and had 82% of prediction of age within a 10 year frame of the true age. The paper also identified the 5 most important markers for predicting human chronological age as albumin, glucose, alkaline phosphatase, urea and erythrocytes.

MRI

In 2018 Tozer et al investigated the correlation between texture parameters on MRI scans and small vessel disease in their “Textured analysis of T1-weighted and fluid-attenuated inversion recovery images detects abnormalities that correlate with cognitive decline in small vessel disease” paper. The study analysis included 118 patients with the mean age of 70 and one texture parameter was found to be a significant predictor of dementia. Another paper looking at signs of ageing from MRI scans is the 2020 Lagner et al. paper “Identifying Morphological Indicators of Ageing with NN on large-scale whole body MRI”. The authors conducted an investigation into age-related changes in whole body MRIs of 32,000 subjects by training a convolutional NN based on VGG16 architecture to predict the age of a given subject based on image data from the scans. They were able to achieve age prediction with mean absolute error of 2.49 years and R^2=0.83. As an input for the network, water and fat signals were compressed by mean intensity projection along the coronal and sagittal directions. The projections were normalised, concatenated and downsampled to 256x256 pixels.

DNA Methylation

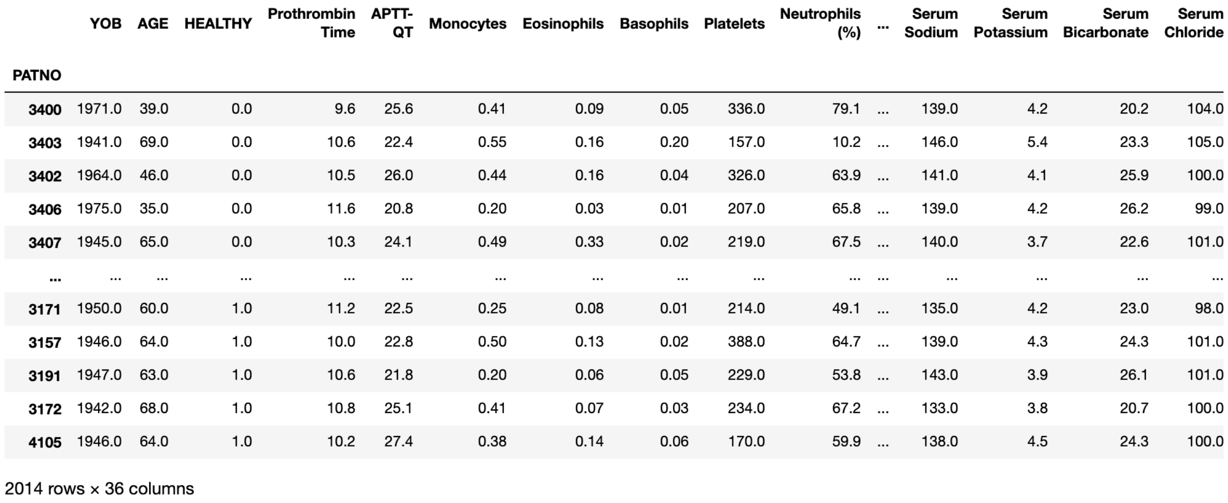
DNA methylation refers to a phenomenon where methyl groups are attached to various sites in an organism’s DNA over the course of its lifetime. This may occur differently in different tissues, and the methylation status of a cell can have important effects on gene expression. In the last several years 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 —out of 27,000 potential sites— 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 3.6 years. Similar results were also obtained by Hannum et al (2014). Working with 656 samples from whole blood, they also used an elastic-net model to identify 71 age-related methylation sites, which they then used them 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, 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..

More recently efforts have been aimed at relating DNA methylation, not to chronological age, but to what is termed biological or phenotypic age as judged by time to death. Levine et al (2018) proposed a model based on 513 methylation sites whose methylation appears to correlate better with time to death than does actual chronological age. And Lu et al (2019) identified DNA methylation sites related to the concentrations of various blood proteins and the degree to which an individual smokes and then used these surrogate DNA methylation sites to also predict mortality. This resulted in an estimate of biological age that they termed DNAm GrimAge that was also a better predictor of mortality than chronological age. The degree to which these two sets of predictors are aligned, however, those for age and those for mortality, is as of yet unclear.

Exploratory Data Analysis and Developed Models

Blood Chemistry

After finding several papers that suggested that blood markers can be predictive of chronological age (see literature review below) we conducted EDA on the PPMI blood chemistry data. The data is in a .csv file containing 217,512 rows corresponding to the results of 33 blood tests for 2014 patients. We reformatted this into a 2014x36 dataframe (see Fig 1 below) adding the year of birth, age at the time of the blood test and patient status information from the Patient_Status.csv file.



	YOB	AGE	HEALTHY	Prothrombin Time	APTT-QT	Monocytes	Eosinophils	Basophils	Platelets	Neutrophils (%)	...	Serum Sodium	Serum Potassium	Serum Bicarbonate	Serum Chloride
PATNO															
3400	1971.0	39.0	0.0	9.6	25.6	0.41	0.09	0.05	336.0	79.1	...	139.0	4.2	20.2	104.0
3403	1941.0	69.0	0.0	10.6	22.4	0.55	0.16	0.20	157.0	10.2	...	146.0	5.4	23.3	105.0
3402	1964.0	46.0	0.0	10.5	26.0	0.44	0.16	0.04	326.0	63.9	...	141.0	4.1	25.9	100.0
3406	1975.0	35.0	0.0	11.6	20.8	0.20	0.03	0.01	207.0	65.8	...	139.0	4.2	26.2	99.0
3407	1945.0	65.0	0.0	10.3	24.1	0.49	0.33	0.02	219.0	67.5	...	140.0	3.7	22.6	101.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3171	1950.0	60.0	1.0	11.2	22.5	0.25	0.08	0.01	214.0	49.1	...	135.0	4.2	23.0	98.0
3157	1946.0	64.0	1.0	10.0	22.8	0.50	0.13	0.02	388.0	64.7	...	139.0	4.3	24.3	101.0
3191	1947.0	63.0	1.0	10.6	21.8	0.20	0.06	0.05	229.0	53.8	...	143.0	3.9	26.1	101.0
3172	1942.0	68.0	1.0	10.8	25.1	0.41	0.07	0.03	234.0	67.2	...	133.0	3.8	20.7	100.0
4105	1946.0	64.0	1.0	10.2	27.4	0.38	0.14	0.06	170.0	59.9	...	138.0	4.5	24.3	100.0

2014 rows × 36 columns

Fig 1: Reformatted blood chemistry data with patient ages, year of birth and status.

Of the 2014 patients with blood chemistry data 232 are healthy controls. The histogram in Fig 2. shows the distribution of the ages, which have a wide enough range for the purposes of exploring the correlation with age.

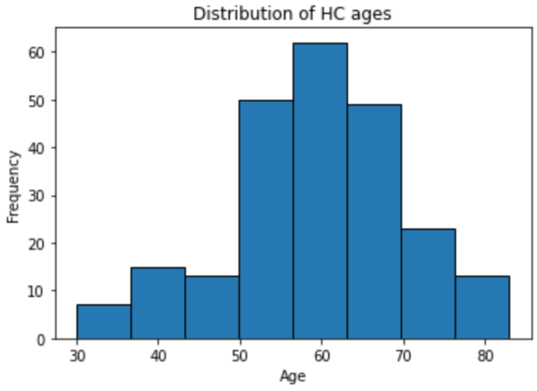
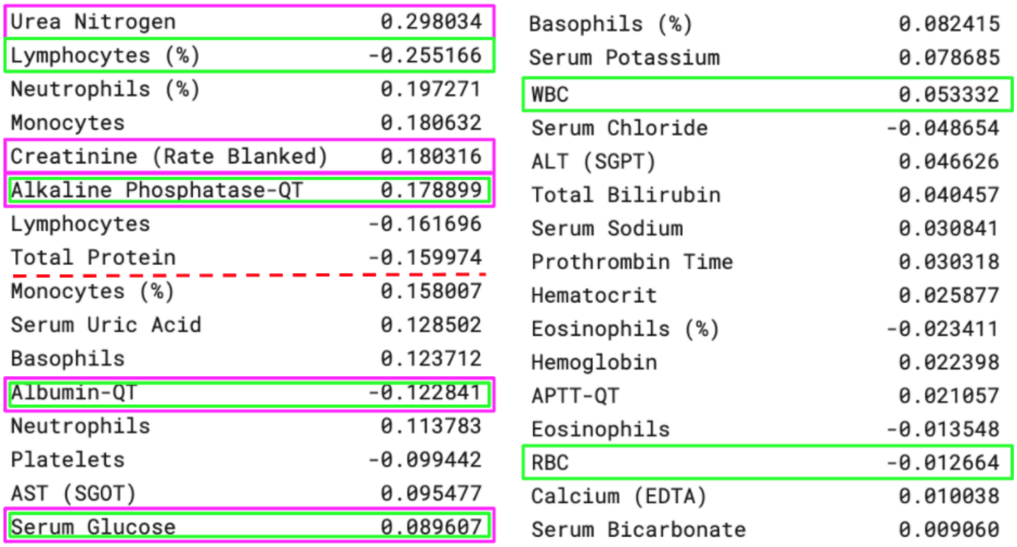


Fig 2: Distribution of the ages of the Healthy Control (HC) patient with blood blood chemistry results

We first looked into the correlation between the blood markers and age for the HC patients. These correlations are listed in order of magnitude in Fig. 3. In the table the markers identified by the Putin study as the best predictors of age are highlighted in pink, and those identified by the Levine study in green. From the figure you can see that individually none of the blood markers appear to have a strong linear correlation with age. The strongest correlation is with Urea Nitrogen with a coefficient of 0.298 followed by Lymphocytes with a coefficient of -0.255. Urea Nitrogen is one of the 5 strongest predictors of age identified by the Putin study and Lymphocytes % is one of the features identified by the Levine study. That being said, these markers still have a relatively weak correlation with age. Interestingly, the three features which were identified by both studies as significant, namely Phosphatase-QT, Albumin-QT and Serum Glucose all have a weak correlation with age in this case with values of 0.179, -0.123 and 0.090 respectively.

Correlation of blood chemistry results with age



```
Urea Nitrogen                 0.298034
Lymphocytes (%)              -0.255166
Neutrophils (%)               0.197271
Monocytes                     0.180632
Creatinine (Rate Blanked)     0.180316
Alkaline Phosphatase-QT       0.178899
Lymphocytes                  -0.161696
Total Protein                -0.159974
Monocytes (%)                 0.158007
Serum Uric Acid               0.128502
Basophils                     0.123712
Albumin-QT                   -0.122841
Neutrophils                   0.113783
Platelets                    -0.099442
AST (SGOT)                    0.095477
Serum Glucose                 0.089607
Basophils (%)                 0.082415
Serum Potassium               0.078685
WBC                           0.053332
Serum Chloride               -0.048654
ALT (SGPT)                    0.046626
Total Bilirubin               0.040457
Serum Sodium                  0.030841
Prothrombin Time              0.030318
Hematocrit                    0.025877
Eosinophils (%)              -0.023411
Hemoglobin                    0.022398
APTT-QT                       0.021057
Eosinophils                  -0.013548
RBC                          -0.012664
Calcium (EDTA)                0.010038
Serum Bicarbonate             0.009060
```

Fig 3: Pearson's correlation coefficients of blood chemistry results with age for the HC group. Pink highlights tests found to be most significant by the Putin study, green highlights tests found to be most significant by the Levine study.

In order to investigate whether there is a relationship between a combination of these features and age, we produced a logistic regression model using the 8 most highly correlated features (listed above the red line in Fig. 3). With a 75-25% train-test split and including a bias term the model performed poorly with an accuracy score of only 0.037. Fig 4 shows the predicted ages against the true ages for the test set

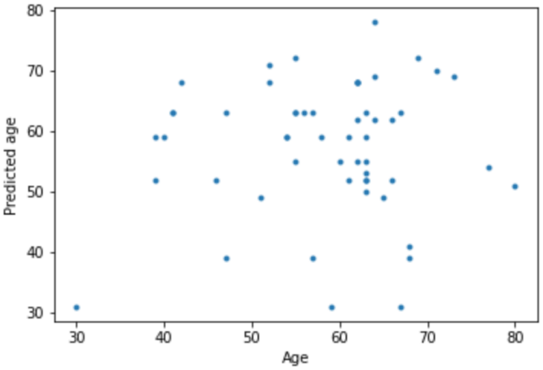


Fig 4: Comparison of the Linear Regression test set predictions vs. the true values

Finally, we produced scatter plots for the 8 most highly correlated features to determine if there's a clear non-linear relationship between these and age. The plots can be seen in Fig 5, however there does not appear to be a clear non-linear relationship. A more complex model, for instance DNN would need to be used to explore the relationship between blood chemistry results and age further.

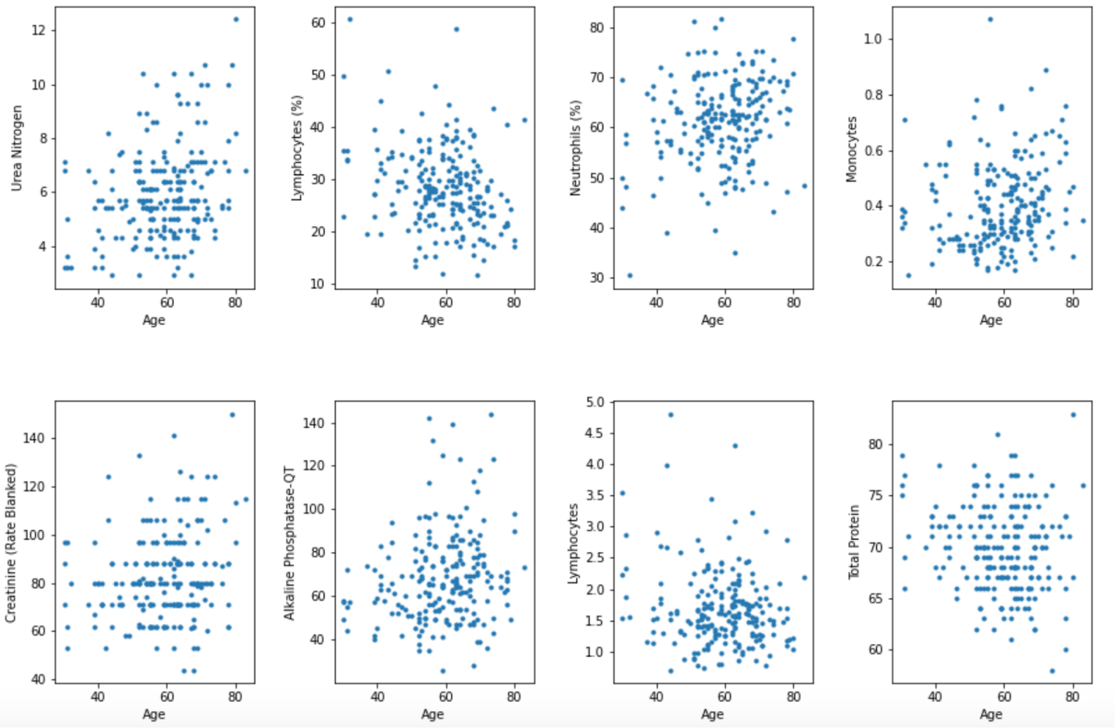


Fig. 5: Scatter plots of the 8 most highly correlated blood test results

After having unpromising correlation results from within the healthy group, we tried to take a step back and see if there is even a true difference between the healthy and unhealthy Parkinson patients. The logic behind that is, if no true difference exists, then we can safely conclude that blood test data is not a great predictor of age, and we won't be able to extract a signal for specifically healthy aging, even if we dive deeper with more sophisticated models and techniques.

Fig 6. shown below shows a distribution of blood test results for different chemical levels. From the violin plot distribution, all metrics have little to no significant difference between the healthy, and the unhealthy samples. We've also looked more closely into chemicals which had an absolute value of more than 0.15 above the average, which also had little to no difference in distribution between the two groups. Even the minimal differences we observed, may likely be due to the underlying age distribution difference between the two groups. Thus, we've concluded after our initial analysis that blood chemical results for

Parkinson’s dataset (or likely neurodegenerative disease) aren’t particularly promising.

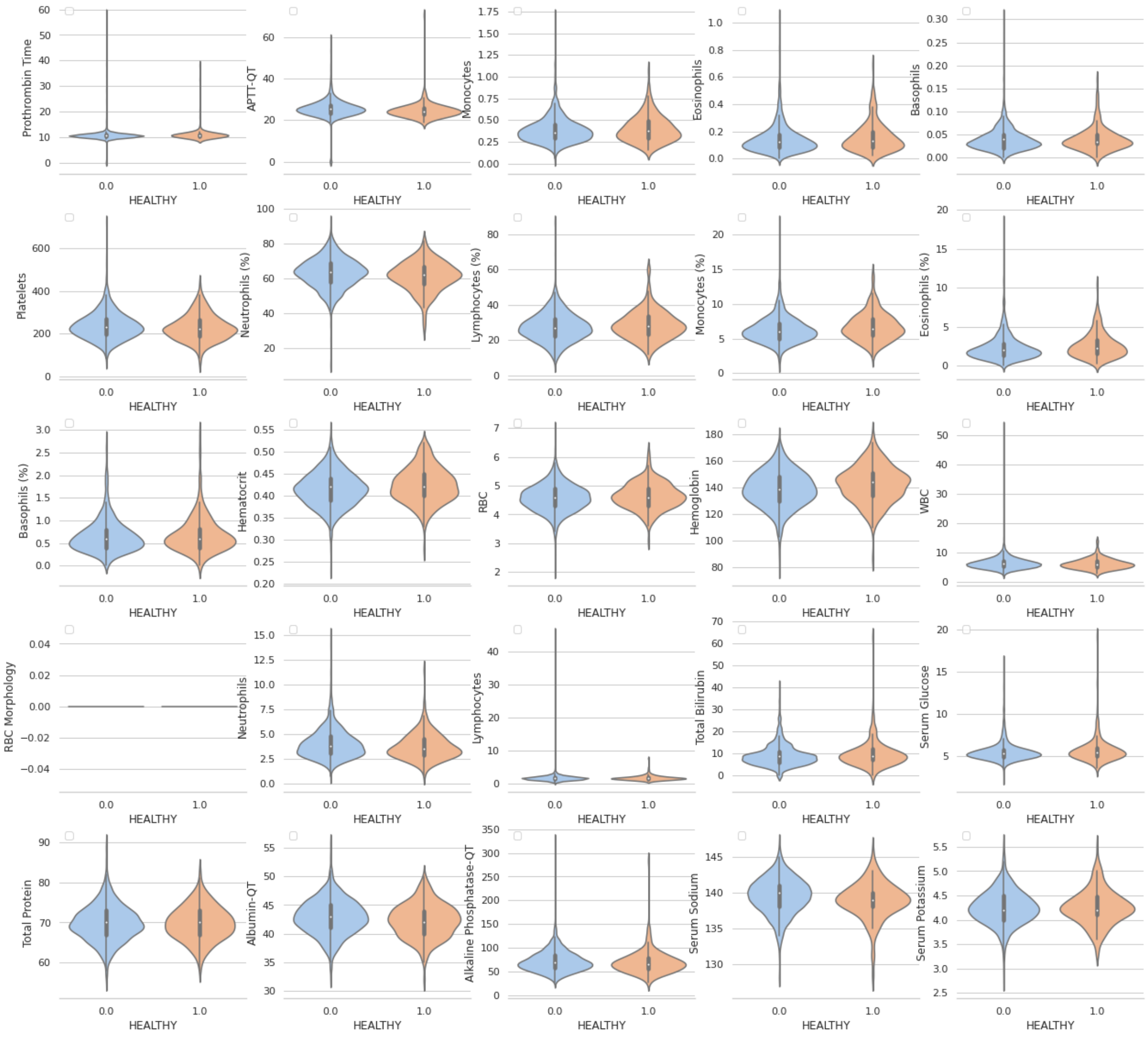
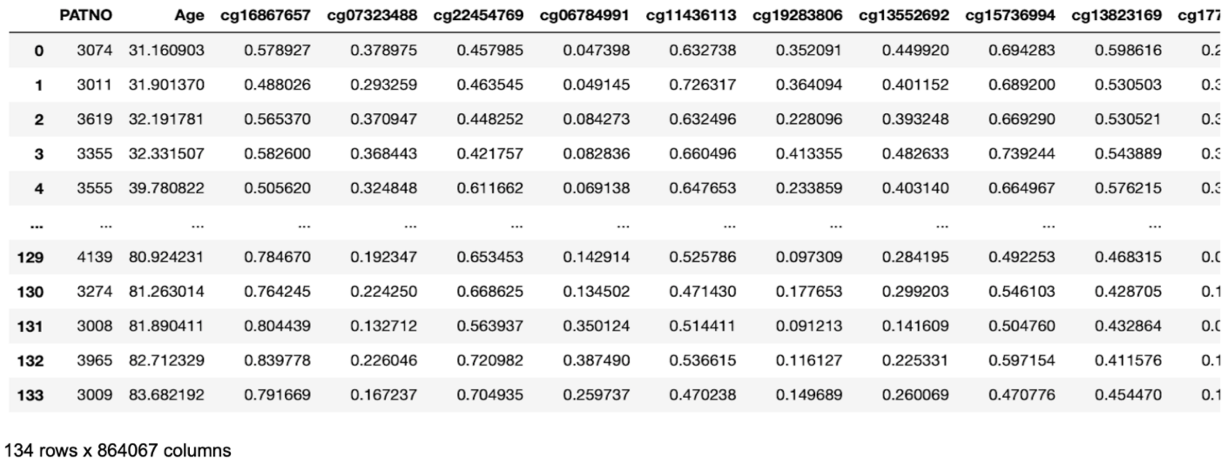


Fig. 6: Distribution of blood test results for Unhealthy samples vs Healthy samples

DNA Methylation

Since there is literature that suggests that DNA methylation correlates with age (see literature review below), in the past week we have examined whether we can see evidence of this in the PPMI DNA-methylation data. As subjects entered the study, their blood was drawn. DNA was isolated from their blood cells, and a DNA-methylation assay was performed. This assay consisted of determining with an illumina infidium microarray, which of 864,067 potential methylations sites, known as cg sites, are in fact methylated for each individual. The data are reported as beta values, that is probabilities of a given site being methylated for each cg site. In the PPMI there is baseline data for 535 subjects. Of these 134 are healthy controls. Thus, the useful data consists of a table of values ranging between 0 and 1 for 134 subjects and 867,067 cg sites. We downloaded these data and transformed them to Table I below.



	PATNO	Age	cg16867657	cg07323488	cg22454769	cg06784991	cg11436113	cg19283806	cg13552692	cg15736994	cg13823169	cg177
0	3074	31.160903	0.578927	0.378975	0.457985	0.047398	0.632738	0.352091	0.449920	0.694283	0.598616	0.2
1	3011	31.901370	0.488026	0.293259	0.463545	0.049145	0.726317	0.364094	0.401152	0.689200	0.530503	0.3
2	3619	32.191781	0.565370	0.370947	0.448252	0.084273	0.632496	0.228096	0.393248	0.669290	0.530521	0.3
3	3355	32.331507	0.582600	0.368443	0.421757	0.082836	0.660496	0.413355	0.482633	0.739244	0.543889	0.3
4	3555	39.780822	0.505620	0.324848	0.611662	0.069138	0.647653	0.233859	0.403140	0.664967	0.576215	0.3
...	...	...	...	...	...	...	...	...	...	...	...	
129	4139	80.924231	0.784670	0.192347	0.653453	0.142914	0.525786	0.097309	0.284195	0.492253	0.468315	0.0
130	3274	81.263014	0.764245	0.224250	0.668625	0.134502	0.471430	0.177653	0.299203	0.546103	0.428705	0.1
131	3008	81.890411	0.804439	0.132712	0.563937	0.350124	0.514411	0.091213	0.141609	0.504760	0.432864	0.0
132	3965	82.712329	0.839778	0.226046	0.720982	0.387490	0.536615	0.116127	0.225331	0.597154	0.411576	0.1
133	3009	83.682192	0.791669	0.167237	0.704935	0.259737	0.470238	0.149689	0.260069	0.470776	0.454470	0.1

134 rows x 864067 columns

Table I. PPMI Methylation Data

Because we have so many features in this dataset, we thought it would be wise to first try to determine which cg sites are most relevant to age, so we isolated those sites whose methylation most correlated with age and created a subset of the data with just the top 1000. In the Fig. 7A below is a boxplot relating methylation degree to age for the top cg site, and the resulting correlation coefficients for the top 1000 cgs are plotted on the right . Of note most are negative (Fig. 7B), which suggests declining methylation with age.

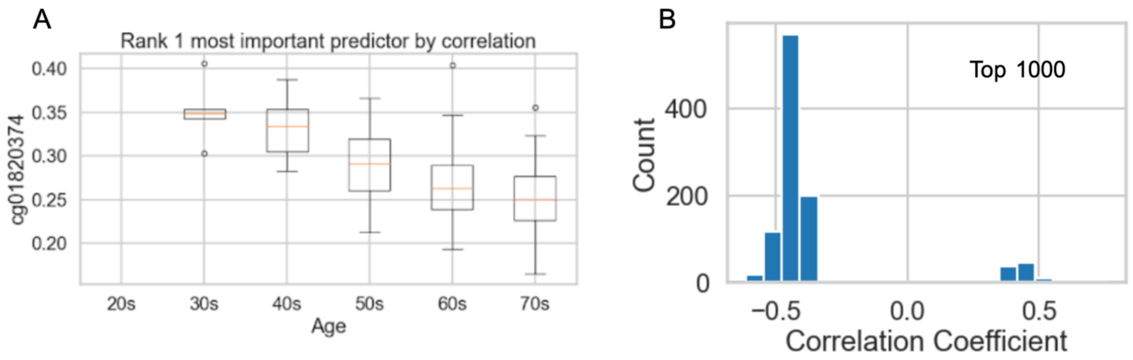


Figure. 7. The most correlated cgs with age

Next, we performed a PCA transformation on these data and plotted the first two principal components of each sample. Encouragingly, as shown in Fig. 8B, there is indeed a separation by age, particularly along the PC1 axis. The lightly colored data points generally lie to the right of the darker data points, and the correlation coefficient between PC1 and age is r = -0.786 (Fig. 8A) Further, when we applied a multiple regression model to these data using the first 10 PCs as features, we were able to account for 84.9% of the variability in age (r^2 = 0.849).

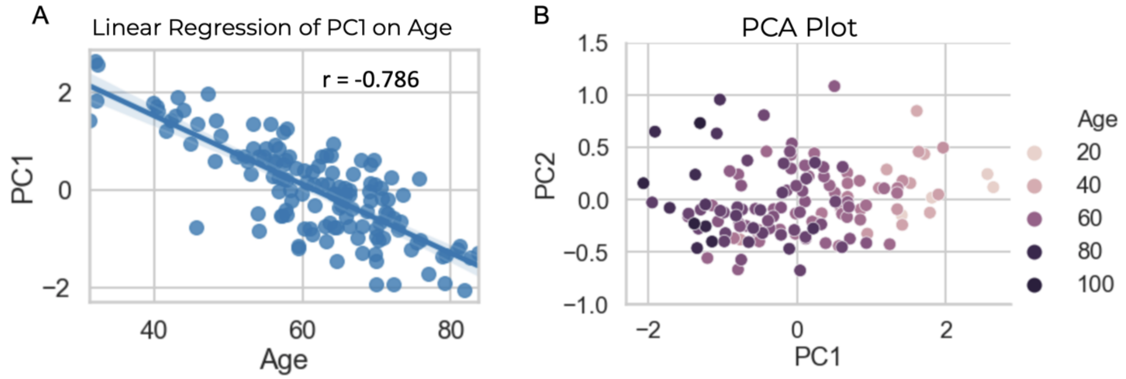


Figure 8. PCA with the top 1000 most correlated cgs

Although the above analysis seemed encouraging, we were concerned about two potential biases. First, by selecting the 1000 cgs out of ~860,000, we may have by random chance selected cgs whose general methylation is not truly correlated with age. To mitigate this issue, we took a second approach to creating a subset of the cg features. We restricted the analysis to just those features that were identified in Horvath 2013 as predictive of age, and we repeated the analysis above. Encouragingly, here again we see a separation on the PCA plot by age (Fig. 9 below), this time more along the PC2 axis. And a multiple regression model with the first 10 PCs yielded an r^2 of 0.729

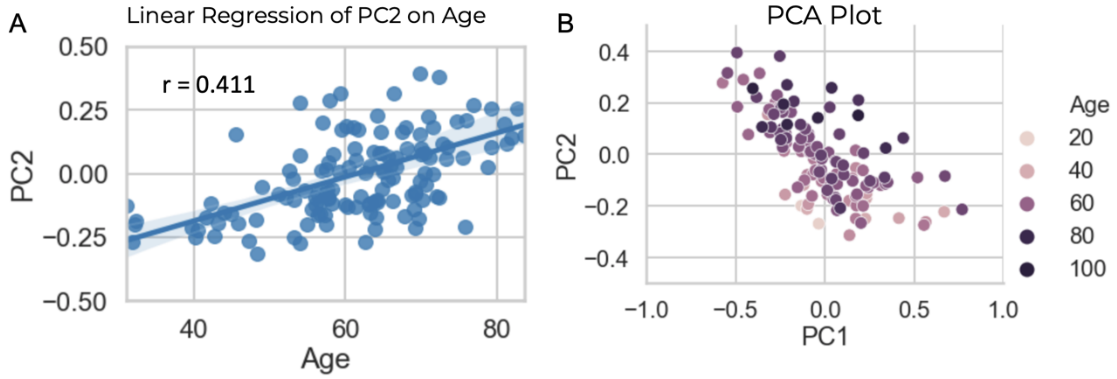


Figure 9. PCA with the 335 cgs identified by Horvath (2013)

Finally, we were also concerned that our regression model may be overfitting the data, so we split the data into train and test sets, 85%-15%, and reran the multiple regression. This analysis yielded a rms on the train set of 8.6 years (r^2 = 0.53) and on the test set of 6.7 years (r^2 = 0.43). Not too bad, however, we found the performance of the model on the test set to be very sensitive to the particular split. The r^2 values ranged from 0.00 to 0.72 across different cross-validation folds, and some cases saw the model perform better on the test set than the training set. This can simply be attributed to the fact that there is not nearly enough data to test the model rigorously. We are now looking for larger data sets. Although these preliminary results are quite unstable, they nevertheless serve as useful baseline models for which future progress can be measured against.

Future Work

Our current plan for future work will mainly focus on the three aspects below:

- **Data:**
  Since our initial EDA shows that blood-test chemical data is not a promising avenue, we've turned our focus to DNA methylation, which performed better during our initial EDA. After deciding our new focus, we realized that the limited size of the PPMI methylation dataset, 200 healthy individuals, may be a disadvantage. Therefore, as mentioned above, to expand our sample size, as well as cover a wider range of ages, we've decided to explore DNA methylation data from the EWAS datahub, which contains over 9000 healthy control samples ranging in ages from infancy to 109 years.. Other datasets we collect during our initial prep process may be taken into account once an accurate baseline model is built and trained on the methylation data we have.

- **Analysis:**
  In our analysis, we define unhealthy narrowly as individuals with neurodegenerative disease e.g. Huntington's, Parkinson's, Alzheimer's, as such a category of disease is also one of the widely agreed diseases to be highly associated with aging.
    - In terms of analysis ideas, we are planning to look at methylation data sampled from different tissues and see if comparing analysis results from different tissues will yield interesting results.
    - Although the focus of our research is on healthy aging, it would be also interesting to do a comparison between healthy and unhealthy groups after we've constructed our model. Such comparison may provide proof that signatures we've included do indeed target healthy samples, or may lead us to new findings.

- **Modeling:** Our current EDA for both methylation data and blood test data uses straightforward naive models. For next steps, we attempt to explore more sophisticated ML models suggested by Merck as well as include high dimensional features, using models such as multi-level perceptron and XGBoost.

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.

Levine, M. E., Lu, A. T., Quach, A., Chen, B. H., Assimes, T. L., Bandinelli, S., et al. (2018). An epigenetic biomarker of aging for lifespan and healthspan. Aging 10, 573–591.

Lu, A. T., Quach, A., Wilson, J. G., Reiner, A. P., Aviv, A., Raj, K., et al. (2019). DNA methylation GrimAge strongly predicts lifespan and healthspan. Aging 11, 303–327. doi: 10.18632/aging.101684

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.