UNIVERSITY OF CALIFORNIA

Los Angeles

Ensemble Based Estimators of a Latent Variable:
Application in Aging Research

A dissertation submitted in partial satisfaction of the
requirements for the degree of Doctor of Public Health

by

Wendy I Ching Shih

2016

ABSTRACT OF THE DISSERTATION

Ensemble Based Estimators of a Latent Variable: Application in Aging Research

by

Wendy I Ching Shih

Doctor of Public Health

University of California, Los Angeles 2016

Professor Stefan Horvath, Chair

Biological age (BA) as opposed to chronological age (CA) is meant to measure the true aging process of an individual. The expectation is that biological age is better than chronological age when it comes to predicting mortality or age related functional decline. We evaluate and adapt techniques from Klemera and Doubal (KD) (2006), an adapted reverse regression approach, and propose a novel ensemble based approach in estimating biological age. However, several simulations has shown that the KD approach may produce unreliable estimates due to the non-robust reverse regression. Though the ensemble based approach mitigates the issue, it is not without limitations. Here, we propose a method to produce more reliable estimates for the KD approach and, consequently, improve the accuracy of our ensemble based approach as well. Lastly, we evaluate and compare the performance of each method including our improved ensemble based model in simulated scenarios and in a Down syndrome application. We found that the ensemble KD approach outperforms the KD approach and standard methods in estimating biological age. Thus, the development of the ensemble KD approach may be useful in examining phenotypic traits that affects aging patterns.

The dissertation of Wendy I Ching Shih is approved.

Christina M Kitchen

David Elashoff

Giovanni Coppola

Stefan Horvath, Committee Chair

University of California, Los Angeles

2016

This doctoral dissertation is dedicated to my wonderful father (Jeff Shih), god-parents (Johnny Sheu and Carol Sheu), sisters (Linda Shih, Lisa Shih, and Katherine Sheu), brother (Victor Sheu), and my forever encouraging friends who cheered for me throughout graduate school and made me believe in the light at the end of the tunnel. Second, I would like to thank my mentors, Dr. Connie Kasari and Dr. Lin Chang, for providing me with so many opportunities, giving me the courage to believe in myself, and allowing me to figure out my niche in statistics. Lastly, I would like to acknowledge the support and guidance of my advisor, Dr. Stefan Horvath, who helped me through every challenge and always reminded me to keep a positive outlook no matter how dismal the situation may seem.

VITA

2011-2016	Statistician, UCLA Semel Institute for Neuroscience & Human Behavior
2010-2016	Statistician, UCLA Oppenheimer Center for Neurobiology of Stress
2009-2014	Teaching Assistant, UCLA Department of Biostatistics
2008-2009	Senior Data Analyst, Kaiser Permanente – Health Plan
2007	M.P.H., Biostatistics, Loma Linda University
2006	B.A. and B.S., Sociology and Statistics, University of California, Los Angeles

PUBLICATIONS (Most recent)

DiStefano, C., **Shih, W.**, Kaiser, A., Landa, R., & Kasari, C. (2016). Communication growth in minimally verbal children with ASD: The importance of interaction. *Autism Research*

Kasari, C., Dean, M., Kretzmann, M., **Shih, W.**, Orlich, F., Whitney, R., ... & King, B. (2016). Children with autism spectrum disorder and social skills groups at school: a randomized trial comparing intervention approach and peer composition. *Journal of Child Psychology and Psychiatry*, *57*(2), 171-179.

Locke, J., **Shih, W.**, Kretzmann, M., & Kasari, C. (2015). Examining playground engagement between elementary school children with and without autism spectrum disorder. *Autism*, 1362361315599468.

Carr, T., **Shih, W.**, Lawton, K., Lord, C., King, B., & Kasari, C. (2015). The relationship between treatment attendance, adherence, and outcome in a caregiver-mediated intervention for low-resourced families of young children with autism spectrum disorder. *Autism*.

Harrop, C., Gulsrud, A., **Shih, W.**, Hovsepyan, L., & Kasari, C. (2015). Characterizing caregiver responses to restricted and repetitive behaviors in toddlers with autism spectrum disorder. *Autism*.

Orand, A., Gupta, A., **Shih, W.**, Presson, A. P., Hammer, C., Niesler, B., ... & Chang, L. (2015). Catecholaminergic Gene Polymorphisms Are Associated with GI Symptoms and Morphological Brain Changes in Irritable Bowel Syndrome.*PloS one*, *10*(8), e0135910.

Chang, Y. C., **Shih, W.**, & Kasari, C. (2015). Friendships in preschool children with autism spectrum disorder: What holds them back, child characteristics or teacher behavior?. *Autism*.

Shire, S. Y., Goods, K., **Shih, W.**, Distefano, C., Kaiser, A., Wright, C., ... & Kasari, C. (2015). Parents' Adoption of Social Communication Intervention Strategies: Families Including Children with Autism Spectrum Disorder Who are Minimally Verbal. *Journal of autism and developmental disorders*, *45*(6), 1712-1724.

Kretzmann, M., **Shih, W.**, & Kasari, C. (2015). Improving peer engagement of children with autism on the school playground: A randomized controlled trial. *Behavior therapy*, *46*(1), 20-28.

Kasari, C., Lawton, K., **Shih, W.**, Barker, T. V., Landa, R., Lord, C., ... & Senturk, D. (2014). Caregiver-mediated intervention for low-resourced preschoolers with autism: an RCT. *Pediatrics*, *134*(1), e72-e79.

Kasari, C., Siller, M., Huynh, L. N., **Shih, W.**, Swanson, M., Hellemann, G. S., & Sugar, C. A. (2014). Randomized controlled trial of parental responsiveness intervention for toddlers at high risk for autism. *Infant Behavior and Development*, *37*(4), 711-721.

Dean, M., Kasari, C., **Shih, W.**, Frankel, F., Whitney, R., Landa, R., ... & Harwood, R. (2014). The peer relationships of girls with ASD at school: comparison to boys and girls with and without ASD. *Journal of Child Psychology and Psychiatry*, *55*(11), 1218-1225.

Shih, W., Patterson, S. Y., & Kasari, C. (2014). Developing an adaptive treatment strategy for peer-related social skills for children with autism spectrum disorders. *Journal of Clinical Child & Adolescent Psychology*, 1-11.

PRESENTATIONS (Most recent)

Shih, W., & Horvath, S. (2015, August). Improving Estimates of Biological Age Using Ensemble-Based Prediction Models in Genomic Data Applications. Oral presentation for Statistics in Epidemiology for Joint Statistical Meeting (JSM), Seattle, Washington.

Chang, Y.C., Shire, S., **Shih, W.**, & Kasari, C. (2015, May). Teacher Implemented Interventions for Preschool Children with Autism: Engagement and Play. Invited guest speaker at UCLA Center for Autism Research and Treatment (CART) conference: Advances in Autism. Los Angeles, California.

Chang, Y.C., Shire, S., **Shih, W.**, Gould, H., & Kasari, C. (2015, May). Diverse Population of Young Children with Autism: Play and Language. Poster presented at the 14th Annual Meeting for Autism Research (IMFAR), Salt Lake City, Utah.

Shih, W., & Shire, S. (2014, May). Getting SMART about Combating Autism with Adaptive Interventions: Novel Treatment and Research Methods for Individualizing Treatment. Symposium panel presentation at the 13th Annual Meeting for Autism Research (IMFAR), Atlanta, Georgia.

Table of Contents

Chapter 1: Introduction

1.1 Introduction to biological age

While aging is a universal phenomenon, the way(s) or rate at which individuals age is not a uniform process, and the method(s) used to define this aging process and determine its association to disease progression are much more debatable. The inability to explicitly define this aging process, however, has not prohibited scientists from creating the concept of a "biological age" as a means to describe this aging process. Biological age is defined as the true global state of aging and is also often referred to as "functional age" and biological age is expected to be better than chronological age at predicting disease status or death.

Biological age is introduced in gerontology research because there is a significant difference between individual subjects with the same chronological age in the rate of age-related changes in certain functions and systems of the human body. Chronological age is an imperfect predictor of mortality and age related functional decline. Biological age, on the other hand, tries to capture an individual's functional status, which can better predict an individual's mortality and functional decline. One thing to note is that biological age is not meant to be replacement for chronological age, but to add additional information that will more accurately represent an individual's health status and longevity potential.

One of the driving forces behind the development of biological age as a concept is to test the possible effectiveness of anti-aging treatments of individuals in longitudinal studies when evaluating effect of environmental toxins on individuals.   A better understanding of biological age would help to illuminate the various environment impacts on the aging process. In addition, better estimates and understanding of biological age are needed to test the effects of early-life adversity, to evaluate social gradients in health, and to search for genetic regulators of aging processes. Differences in biological age among racial groups or disease status could also reflect genetic differences, or gene by environment interactions. Hence, biological age measures may be used as a valuable phenotype for examining genetics of human aging.

Though the concept of biological age has been referenced in many publications, there is little consensus regarding the method in which biological age should be estimated. Since validation of biological age has always been a controversial issue, as the term itself is an abstract concept, it difficult to evaluate the validity of the estimated biological age and absence of the true value makes it impossible to do so in practice. Multiple methods have been proposed to estimate biological age, but no gold standards have been established or are widely mainly agreed upon due to two main reasons: what biomarkers should be used and what type of models or algorithms should be used for estimation. The current popular methods for estimating biological age include but are not limited to: multiple linear regression, principal component analysis, factor analysis, and penalized regression.

1.2 Current widely use methods

1.2.1 Multiple Linear Regression (MLR)

Multiple linear regression (MLR) is the simplest and most common approach in biological age related research (e.g. Hollingsworth et al., 1965; Takeda et al., 1982; Voitenko and Toka, 1983; Dubina et al., 1984; Kroll and Saxtrup, 2000; Bae et al., 2008). In these studies, biological age is estimated by several predictors (e.g. biomarkers) that have been shown to be associated with aging with chronological age as the outcome. In the multiple linear regression method, the predicted chronological age given the biomarkers is assumed to be equal to biological age. The parameters are estimated from the least squares method where the sum of squared residuals is minimized over all possible values of the intercepts and slopes.

The linear model assumes that chronological age is predicted from the measured values of the biomarkers (i.e. independent variables, predictors, traits, or features). However, any change in chronological age does not depend on changes in the biomarkers, but rather on the calendar year. Consequently, many criticisms of using multiple linear regression to predict biological age target this issue and claims that predicted chronological age is not equivalent to biological age.

Other criticisms of utilizing multiple linear regression in estimating biological age involves the thousands of biomarkers that could easily be collected due to the advancement of genetic research. Many of the past studies would pre-define a set of biomarkers that are highly correlated with chronological age as input for the predictors. Researchers would also use stepwise regression to find the best subset of biomarkers using either p-values, AIC, BIC, or other statistical measures to determine the entry/dropout criterion. Using a stepwise regression approach to define the best subset of predictors, however, has its disadvantages which are well known and will not be discussed in great length here. Another issue with using multiple linear regression is multicollinearity. Since many past studies have already predefined a set of biomarkers that are highly correlated with chronological age, the biomarkers are most likely highly correlated with each other which will result in unstable estimates and inflated variances.

1.2.2 Principal Component Analysis (PCA)

Principal component analysis is a common mathematical procedure that uses a transformation to convert a set of observations of possibly correlated variables into a set of independent components to deal with multicollinearity and large data dimension. The transformation is performed such that the first component explains the most variation in the original variables and the second component explains the second most variation and so forth. Many studies utilized principal component analysis to summarize most of the original information into a minimum number of components.

The first principal component score is often used to signify an estimate of biological age. One way in which principal component analysis can be used to predict biological age is to transform the first principal component. Given that the first component is not in units of years, the scores were transformed to allow for comparisons with chronological age. An alternative suggested by Nakamura (1991) is to adjust the transformed component by adding a z-score to the biological age estimates, in order to account for systematic errors that may cause over or under estimations of biological. However, the final estimator of biological age is still multiple linear regression. Hence, the limitations that are prevalent in multiple linear regression are still present even when the components are the new set of predictors.

1.2.3 Factor Analysis (FA)

Factor analysis (FA) is similar to principal component analysis in that it is a technique for examining the interrelationships among a set of variables. Both of these techniques differ from regression analysis in that we do not have a dependent variable to be explained by a set of independent variables. However, principal component analysis and factor analysis also differ from each other. In principal components analysis the major objective is to select a number of components that explain as much of the total variance as possible. The values of the principal components for a given individual are relatively simple to compute and interpret. On the other hand, the factors obtained in factor analysis are selected mainly to explain the interrelationships among the original variables.

In factor analysis, a battery of variables is usually standardized so that their variances are each equal to one and their covariances are the correlation coefficients. The objective of factor analysis is to represent each of these variables as a linear combination of a smaller set of common factors. Once the initial extraction of factors and the factor rotations are performed, it may be of interest to obtain the score an individual has for each factor. A regression procedure is commonly used to compute factor scores where the method combines the intercorrelations among the original variables and the factor loadings to produce factor score coefficients. These coefficients are used in a linear fashion to combine the values of the original standardized variables into factor scores. Similar to a principal component approach, as described in previous section, the final estimation of biological age is still multiple linear regression. Hence, the limitations that are prevalent in multiple linear regression are still present even when the components are the new set of predictors (Affifi et.al 2011).

1.2.4 Penalized Regression (Elastic-net)

The elastic-net regression falls under the overarching branch of penalized regression. Penalized regression is often employed as a means of variable selection. One of the major downfalls of multiple linear regression is the difficulty in interpretation when there are a large number of predictors. When presented with a large number of predictors, we often prefer to determine a smaller subset that exhibits the strongest effects. By retaining a subset of the predictors and discarding the rest, subset selection produces a model that is interpretable and that possibly has a lower prediction error than the full model (Trevor et. al. 2001). Penalized regression shrinks the regression coefficients by imposing a penalty or penalties. Therefore, the elastic-net is a penalized method imposing an L1 and L2 penalty on the regression coefficients. For the sake of simplicity, we can consider L1 and L2 penalties in the least squares application as regularization terms in order to prevent the coefficients to overfit. The difference between the L1 and L2 penalties is just that the L2 is the sum of the square of the weights, while L1 is just the sum of the weights.

Consider a linear regression with a given response values y_i and predictors $x_{ij}$for i=1, 2, . . . ,N where N is the number of samples and j=1, 2, . . . ,p, where p is the number of predictions. A naïve elastic-net estimator $\hat{\beta}$ minimizes

$$\sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2 + \lambda_1 \sum_{j=1}^{p}\left|\beta_j\right| + \lambda_2 \sum_{j=1}^{p} \beta_j^2 ,$$

where $\lambda_1 \sum_{j=1}^{p}\left|\beta_j\right|$is the L1 penalty and $\lambda_2 \sum_{j=1}^{p} \beta_j^2$ is the L2 penalty. The algorithm can be considered as a penalized least squares method where $\alpha = \lambda_2/(\lambda_1 + \lambda_2)$. Solving $\hat{\beta}$ in the previous equation is equivalent to the optimization problem

$$\hat{\beta} = \arg min \sum_{i=1}^{N}\left(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\right)^2$$

with the constraint $(1-\alpha)\sum_{j=1}^{p}\left|\beta_j\right| + \alpha \sum_{j=1}^{p} \beta_j^2 \leq t$ for some *t*. The function

$$(1-\alpha)\sum_{j=1}^{p}\left|\beta_j\right| + \alpha \sum_{j=1}^{p} \beta_j^2$$

is known as the elastic-net penalty which is a convex combination of the lasso and ridge penalty (Zhou et. al. 2005).

1.3 Chapter 1 References

Afifi, A., May, S., & Clark, V. A. (2011). *Practical multivariate analysis*. CRC Press.

Bae, C. Y., Kang, Y. G., Kim, S., Cho, C., Kang, H. C., Yu, B. Y., ... & Kim, J. S. (2008). Development of models for predicting biological age (BA) with physical, biochemical, and hormonal parameters. *Archives of gerontology and geriatrics*, *47*(2), 253-265.

Dubina, T. L., Mints, A. Y., & Zhuk, E. V. (1984). Biological age and its estimation. III. Introduction of a correction to the multiple regression model of biological age and assessment of biological age in cross-sectional and longitudinal studies. *Experimental gerontology*, *19*(2), 133-143.

Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2005). The elements of statistical learning: data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83-85.

Hollingsworth, J. W., Hashizume, A., & Jablon, S. (1965). Correlations between tests of aging in Hiroshima subjects--an attempt to define" physiologic age". *The Yale journal of biology and medicine*, *38*(1), 11.

Hollingsworth, M. J., & Bowler, K. (1966). The decline in ability to withstand high temperature with increase in age in Drosophila subobscura.*Experimental Gerontology*, *1*(4), 251-257.

Krøll, J., & Saxtrup, O. (2000). On the use of regression analysis for the estimation of human biological age. *Biogerontology*, *1*(4), 363-368.

Nakamura, E. (1991). A study on the basic nature of human biological aging processes based upon a hierarchical factor solution of the age-related physiological variables. *Mechanisms of ageing and development*, *60*(2), 153-170.

Takeda, H., Inada, H., Inoue, M., Yoshikawa, H., & Abe, H. (1981). Evaluation of biological age and physical age by multiple regression analysis. *Medical informatics= Medecine et informatique*, *7*(3), 221-227.

Voitenko, V. P., & Tokar, A. V. (1983). The assessment of biological age and sex differences of human aging. *Experimental aging research*, *9*(4), 239-244.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.

Chapter 2: Klemera and Doubal’s Approach

2.1 Method

In their article, “A new approach to the concept and computation of biological age,” Klemera and Doubal (KD) (2006) presented a new mathematical algorithm that they proposed as an optimum method of estimation for biological age. The KD approach shared similar features to the reverse regression approach as Hochchild (1989) who developed a new algorithm in order to avoid the common statistical pitfalls of multiple linear regression. Rather than regressing chronological age on the biomarkers, Klemera and Doubal regressed every individual biomarker on chronological age and extracted the estimated coefficients from each simple linear regression as new weights for their final models instead of regressing chronological age on the biomarkers.

In reviewing many published results of biological age estimates using multiple linear regression, Klemera and Doubal found that chronological age is typically a more precise estimate of biological age than estimates computed by multiple linear regression. They suggest that chronological age should be used as a standard biomarker, thus improving the precision of biological age estimate. The biological age estimates are based on minimizing the distance between m regression lines and m biomarker points, within an m dimensional space of all biomarkers. In their article, the authors used computer-generated simulations to validate the method they proposed.

Klemera and Doubal presented two alternative methods for calculating the optimum estimates of biological age. First, they define the relationships among biological age (BA), chronological age (CA), and the biomarker measurement (X) as:

$$BA = CA + R_{BA}(0, s_{BA}^2)$$

$$X = F_X(BA) + R_X(0, s_X^2)$$

where $R_{BA}(0, s_{BA}^2)$ and $R_X(0, s_X^2)$ are random variables with mean zeroes and variance s_{BA}^2, s_X^2 respectively and F_X is a function of a biomarker by biological age. By combining the two equations, $BA = CA + R_{BA}(0, s_{BA}^2)$ and $X = F_X(BA) + R_X(0, s_X^2)$, and assuming F_X is a linear function with intercept q and slope k, the resulting equation becomes

$$X = kCA + q + R(0, k^2 s_{BA}^2 + s_X^2)$$

and the expected value of biomarker X_j is $\mu(X_j) = k_j t + q_j$ where *t* represents the biological age. Hence, this implies that the expected value of X_j is equivalent to a function of t:

$$\mu(X_j) \approx F_j(t).$$

Consequently, the distance between the expected value of the biomarker and its measured value can be measured as a weighted sum of squares which is just another function of time *t* and is denoted as $Q(t)$,

$$Q(t) = \sum_{j=1}^{p} \alpha_j \left[F_j(t) - x_j\right]^2$$

where α_j is the individual weights for biomarker X_j. Klemera and Doubal found the optimal choice of α_j to be $1/s_j^2$ where s_j^2 represents the mean squared error from the regression model regressing jth biomarker on biological age. However, given that the biological age is not measurable or observable, the mean squared errors from the regressions between each biomarker and chronological age can be used as replacements as suggested by Cho, Park, and Lim (2009). In order to determine the optimum estimate of biological age, Klemera and Doubal solved for the solution of the equation $Q'(t) = 0$ for the unknown *t* resulting in the two equations, equations (1) and (2), in which the later method utilizes chronological age in the final equation.

$$BA_E = \frac{\sum_{j=1}^{m}(x_j - q_j)\left(\frac{k_j}{s_j^2}\right)}{\sum_{j=1}^{m}\left(\frac{k_j}{s_j}\right)^2} \qquad (1)$$

$$BA_E = \frac{\sum_{j=1}^{m}(x_j - q_j)\frac{k_j}{s_j^2} + \frac{C}{s_B^2}}{\sum_{j=1}^{m}\left(\frac{k_j}{s_j}\right)^2 + \frac{1}{s_B^2}} \qquad (2)$$

where x_j corresponds to the j^th^ biomarker/covariate, q_j is the intercept from the regression regressing chronological age on the j^th^ biomarker, and k_j is the slope of chronological age from the regression regressing chronological age on the j^th^ biomarker. Using simulations, Klemera and Doubal showed that equation (2) to be superior by having smaller error.

In order to produce an estimate for biological age, s_j^2 and s_B^2 need to be estimated. As mentioned previously, s_j^2 are derived from the mean squared errors from the regressions between each biomarker and chronological age. In order to calculate s_B^2, Klemera and Doubal used a characteristic value r_{char} of a group of m various correlation coefficients r_j and sequentially computed s_B^2.

$$r_{char} = \frac{\sum_{j=1}^{m}\frac{r_j^2}{\sqrt{1-r_j^2}}}{\sum_{j=1}^{m}\frac{r_j}{\sqrt{1-r_j^2}}}$$

$$s_B^2 = \left(\frac{\sum_{j=1}^{n}\left((BA_i - CA_i) - \sum_{i=1}^{n}(BA_i - CA_i)/n\right)^2}{n}\right) - \left(\frac{1-r_{char}^2}{r_{char}^2}\right) \times \left(\frac{(CA_{max} - CA_{min})^2}{12m}\right)$$

The value r_j^2 used to calculate the characteristic correlation coefficient r_{char} correlation coefficient between chronological age and the *m* biomarkers.

2.2 Applications of Klemera and Doubal’s Approach in Literature

The KD approach has been applied in several aging studies in literature since its introduction. Though many articles have referenced Klemera and Doubal’s study, only Cho et al. (2009), Levine (2013), Levine and Crimmins (2014), Schaefer et al. (2015), and Belsky et al. (2015) have applied the KD approach in estimating biological age in their studies to the best of our understanding. Among the five studies, only Cho et al (2009), and Levine (2013) compared the KD approaches with other methods of estimating biological age; whereas, the remaining three studies solely applied the KD approaches in estimating biological age in their studies.

2.2.1 Cho, Park, and Lim (2009)

Cho et al. (2009) compared the estimated biological ages from the KD approaches with the estimated biological age from multiple linear regression, principal component analysis and Hochild’s reverse regression. The purpose of their study was to find the optimal method of estimating biological age by examining the association between the estimated biological age with the Work Ability Index (WAI) and comparing the differences of each algorithm's estimates from chronological age as measurements of error. The WAI was found to function as a measure that reflects an individual's current health status rather than the deterioration caused by a serious dependency with the age. Hence, Cho et al. considered the WAI to be the golden standard with which to compare their estimates.

The data were collected from 200 Korean male participants and measurements include biomarkers related to physical function (i.e. hearing capacity, pulmonary functions, handgrip strength, vibrotactile sensitivity, visual accommodations), cognitive functions (i.e. numeric memory, associated memory, topological memory, and concentration), and reaction times (i.e. acoustic reaction time, visual reaction time, and muscular reaction time). Cho et al. found that the estimated biological ages estimated from the KD approaches were the most correlated with WAI, indicating that the estimates of the methods adequately correspond to the health status of the individuals. Cho et al. also noted that although estimates from KD1 showed better correlation with WAI than estimates from KD2, they suggested the usage of KD2 since KD1 might suffer from violation of the assumptions that all biomarker variables should be uncorrelated and that the validity of the KD approaches depends on meeting such an assumption.

2.2.2 Levine (2013)

In her 2013 study, Levine aimed to compare the estimated biological age estimated from multiple linear regression, principal component analysis, and the KD approaches with the objective of determining their validity and usefulness in predicting mortality outcomes, within a large nationally representative human sample. The author compared the predictive ability of these algorithms that estimate biological age by using a sample that included 9,389 persons, aged 30–75 years from the National Health and Nutrition Examination Survey (NHAHES) III. The biomarkers chosen were based on the author’s knowledge regarding their role on the aging process or usage in previous biomarkers of aging studies, their availability in the NHANES data set, and their statistical association with chronological age. The final ten biomarkers selected are C-reactive protein, serum creatinine, glycated hemoglobin, systolic blood pressure, serum albumin, total cholesterol, cytomegalovirus optical density, serum alkaline phosphatase, forced expiratory volume, and serum urea nitrogen.

The KD approaches produced estimates of biological age that performed the best with the highest sensitivity with chronological age, particularly when the predictors are chosen by principal component analysis. The KD2 produced reasonable estimates of biological age and was a more consistent predictor of mortality than chronological age or any of the other biological age estimated from multiple linear regression and principal component analysis algorithms in multiple age cohorts. Lastly, the author reported that when included with chronological age in a model, the KD estimates had more robust predictive ability resulting in chronological age to no longer be significantly associated with mortality. Levine concluded that given the KD approaches' ability to use a single measure to combine a number of varying biomarkers, the KD approaches accounted for the complexity of aging in its measurement.

2.2.3 Levine and Crimmins (2014)

Racial disparities are associated with astonishing health disparities in the United States especially among the African American population. African American experience morbidity and mortality earlier in the life course compared to Caucasians which may be an indicator of accelerated aging. Following her 2013 study, Levine and Crimmins (2014) examined the racial difference in the rate of aging across ages groups to determine if African American aged biologically faster than Caucasians and whether disparities in the aging patterns decline in later years. The authors utilized data from the NHANES III between 1988 and 1994 and they applied the KD2 estimate as an estimation of biological age. The same ten predictors as the Levine (2013) study were utilized in the KD2 estimate to predict biological age.

Levine and Crimmins found that African Americans were significantly higher than Caucasians by 3 years after adjusting for chronological age and sex. Even after adjusting for socioeconomic status and health behaviors, African Americans were still biologically older than Caucasians. One pertinent finding from this study was the attenuation in aging disparities between African American and Caucasian after the age of 70. African American in their 30's, to 60's had biological ages that were 2.28, 3.63, 4.59, and 4.82 years, respectively, higher than Caucasians. However, the age disparities started to decrease after the age of 70 where the racial differences in biological age decreased to 2.94 and 1.17, respectively, and were no longer significant different for people post 80's.

2.2.4 Belsky et al. (2015)

Belsky et al. conducted a very interesting study where they studied aging in a population-representative 1972–1973 birth cohort of 1,037 young adults followed from age 28 to age 38 from the Dunedin Study. The authors utilized the same estimation from Levine (2013) where they utilized the KD2 algorithm with the ten biomarkers that Levine (2013) suggested. Since the Dunedin Study was a longitudinal study, the authors were also able to calculate a "pace of aging" by capturing within-individual longitudinal change in 18 biomarkers across ages 26, 32, and 38 as a measurement of each study member's personal rate of physiological deterioration. This pace of aging is similar in concept as the "rate of aging" referenced in later text. The pace of aging was calculated in three steps: 1) standardization of each of the 18 biomarkers, 2) calculating individual slopes for each of the 18 standardized biomarkers using a mixed effects growth model with random intercepts and random slopes that regressed the biomarker measurement on chronological age, and 3) summing the random slopes of all 18 biomarkers for each subject.

The authors found that subjects who were biological older at 38 years displayed an accelerated pace of aging from age 26-38, performed less optimal on objective tests of physical functioning at age 38 than biologically younger peers, had more difficulty with balance and motor tests, had weaker grip strength, had poorer cognitive functioning at midlife, and showed decline in cognitive performance net of their baseline level. The results showed that the aging process can be readily quantifiable in young adults where age-related diseases had not started to emerge. In addition, the study also provided description and evidence between the differences between the pace of aging and biological age. Pace of aging and biological age are two different approaches to quantifying aging. The authors showed that biological age can be used to provide a summary of accumulated aging in cases where only cross-sectional data are available. However, for purposes of measuring the effects of risk exposures and antiaging treatments on the aging process, pace of aging measures are more suitable to provide a means to evaluate within individual change.

2.2.5 Schaefer et al. (2015)

Different from Cho et al. (2009) and Levine (2013) studies, Schaefer et al. applied several estimation of biological ages in order to examine whether intelligence could predict measures of aging at midlife before the onset of most age-related diseases and not to compare the validity and reliability of each biological age estimates. Interestingly, Schaefer et al. utilized perceived facial age, Framingham heart age, mean relative leukocyte telomere length, and estimates from KD2 using the ten biomarkers identified from Levine (2013) as their estimates of biological age. Intelligence was measured at early childhood (ages three to five), middle childhood (ages seven, nine, and eleven), and at midlife (age 38) using the Peabody Picture Vocabulary (PPVT), Reynell Developmental Language Scores (RDLS), Wechsler Intelligence Scale (WISC-R). Participants of the study were from the Dunedin Study born between 1972 and 1973.

Schaefer et al. found that participants with higher biological ages (perceived facial age, KD approach, and Framingham heart age) at midlife were significantly associated with lower intelligence at early childhood, middle childhood, and midlife. However, the strength of the associations was mostly weak (r range: 0.09 – 0.189). Surprisingly, telomere length was not significantly associated with intelligence at all age groups expect for middle childhood. However, the association between telomere length and middle childhood intelligence was relatively weak (r=0.073), and most likely, the significance was driven by the large sample size (N=1,073). The authors concluded that early-life cognitive enhancement interventions may help to decrease or delay age-related morbidity.

2.3 Conclusion: Why use KD Approach as the individual model of the ensemble method?

As mentioned in Chapter 1, there are currently no standard algorithms for computing biological age and a true biomarker of aging is yet to be defined. With a multitude of algorithms to choose from, we chose the Klemera and Doubal approach because of several reasons mentioned in Chapter 1 and the applications described in this chapter. As Klemera and Doubal pointed out in their 2009 publication, "With respect to these facts parameters of many published batteries of BMs, where MLR-method was used, show that the estimated BAs have greater error than possible standard deviation of differences BA from CA might be." Hence, using the chronological age may be a better estimate of true biological age than the estimated biological age estimated from multiple linear regression. Hence, due to this reason, we are hesitant from using linear regression approaches in choosing our individual model.

Another reason for choosing the KD approaches was attributed to studies conducted by Cho et al. (2009) and Levine (2013). Both studies compared the KD approaches with standard methods such as multiple linear regression and principal component analysis and both studies found that the estimates from the KD approaches to be superior. This lends support to the usage of the KD approaches in estimating biological age over standard estimating methods. Lastly, the studies conducted by Levine and Crimmins (2014), Belsky et al., (2015), and Schaefer et al. (2015) showed how the KD approaches can be used in various way in nationally large representative studies (NHANES and Dunedin Study) which indicate the versatility of the KD approaches.

Though the current study is to apply and adapt from the KD approaches in aging research, we are not discrediting the value of standard methods in estimating biological age. Multiple linear regression has its advantage with well-established diagnostic tools such as examining the residual plots and diagnostic statistics such as Cook's distance measure to detect any potential outliers. Furthermore, linear regression also hosts a multitude of great diagnostic tools for multicollinearity which can be a great asset in constructing a valid set of biomarker variables for biological age estimation. Principal component analysis also has displayed its worth in biological age estimation. Not only it generates uncorrelated variables, but it also provides information about the underlying structure of the variables. Hence, many studies in biological aging utilizes principal component analysis to summarize a battery of biomarkers and then uses multiple linear regression for estimation. Cho et al. (2009) also noted that utilizing the underlying structure of the biomarkers identified from PCA provided them a means to make the additional application of the KD approaches. Consequently, the usage of standard methods in biological age estimation is still quite practical in many settings.

2.4 Chapter 2 References

Belsky, D. W., Caspi, A., Houts, R., Cohen, H. J., Corcoran, D. L., Danese, A., ... & Sugden, K. (2015). Quantification of biological aging in young adults. *Proceedings of the National Academy of Sciences*, *112*(30), E4104-E4110.

Cho, I. H., Park, K. S., & Lim, C. J. (2010). An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI). *Mechanisms of ageing and development*, *131*(2), 69-78.

Hochschild, R. (1989). Improving the precision of biological age determinations. Part 1: A new approach to calculating biological age.*Experimental gerontology*, *24*(4), 289-300.

Klemera, P., & Doubal, S. (2006). A new approach to the concept and computation of biological age. *Mechanisms of ageing and development*,*127*(3), 240-248.

Levine, M. E. (2013). Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age?. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *68*(6), 667-674.

Levine, M. E., & Crimmins, E. M. (2014). Evidence of accelerated aging among African Americans and its implications for mortality. *Social Science & Medicine*, *118*, 27-32.

Schaefer, J. D., Caspi, A., Belsky, D. W., Harrington, H., Houts, R., Israel, S., ... & Moffitt, T. E. (2015). Early-life intelligence predicts midlife biological age. *The Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, gbv035.

Chapter 3: Epigenetic Clock

3.1 Introduction to the Epigenetic Clock

Many of the techniques used in validation and methods described in this study mirrors the approaches described in Hovarth's 2013 study (Horvath, 2013). Horvath developed a multi-tissue predictor of age that allows one to estimate the DNA methylation age (DNAm age) of most tissues and cell types. The predictor, which is freely available in R, was developed using 8,000 samples from 82 Illumina DNA methylation array datasets, encompassing 51 healthy tissues and cell types. Lastly, Horvath characterized the 353 CpG sites that together form an aging clock in terms of chromatin states and tissue variance.

3.2 Methods

Based on the training set data, Horvath found that it is advantageous to transform age before carrying out an elastic-net regression analysis. Detailed information on elastic-net regression is described in Section 1.2.4. Horvath also used the following function F for transforming chronological age (CA) prior to estimation:

$$F(CA) = \begin{cases} log(CA+1) - log(adult.age+1) & if\ CA \leq adult.age \\ (CA - adult.age)/adult.age + 1 & if\ CA > adult.age \end{cases}$$

The constant, adult.age, was set to age 20 for humans and age 15 for chimpanzees and the function *F* satisfies the following desirable properties:

1) continuous, monotonically increasing function (which can be inverted),

2) logarithmic dependence on age until adulthood (here set at 20 years),

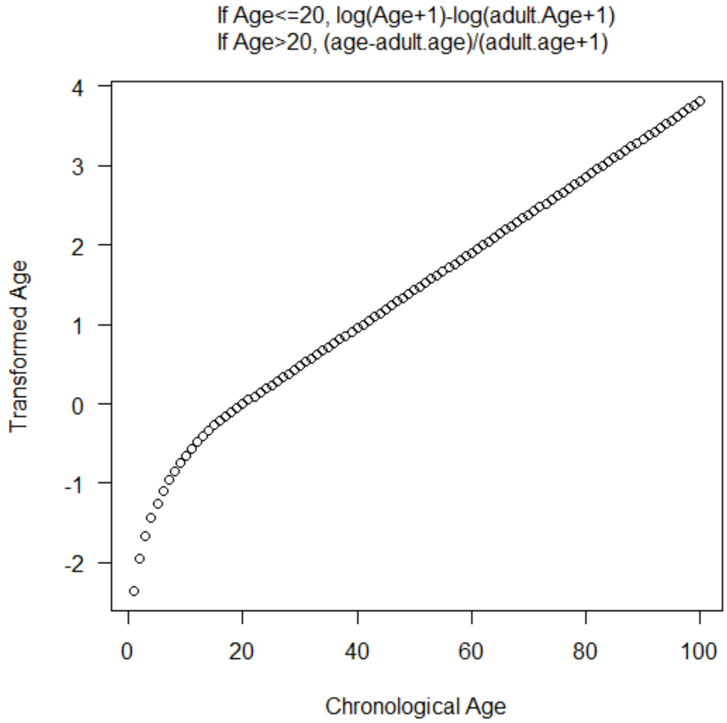
3) linear dependence on age after adulthood (here set to 20),

4) defined for negative ages (i.e. prenatal samples) by adding 1 (year) to age in the logarithm,

5) continuous first derivative (slope function). In particular, the slope at age=adult.age is given by 1/(adult.age+1).

The function *F* is visualized in Figure 3.1 which shows the association between the chronological age and the transformed chronological age. Prior to the age of 20, subjects experience an accelerated aging and aging starts to slow down to a steady increasing rate post 20 years old.

Figure 3.1: Transformation of Chronological Age



The inverse of the function F, denoted by inverse.F, is used to transform the linear part of the regression model into DNAm age. An elastic net regression model (implemented using the "*glmnet*" R function) was used to regress a transformed version of age on the roughly 21k beta values in the training data. The elastic net regression resulted in a linear regression model where coefficients β_0, β_1, ..., β_{353} relate to transformed age as follows

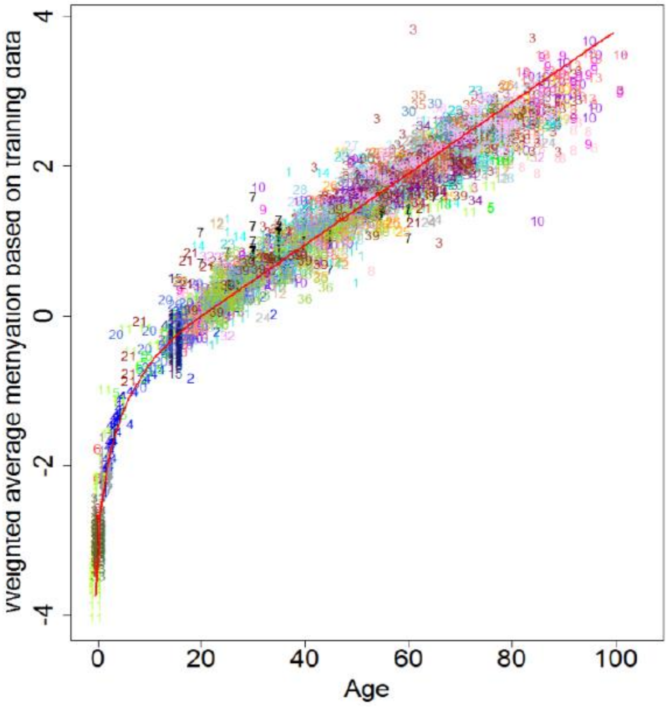
$F(chronological\ age) = \beta_0 + \beta_1 CpG_1 + \cdots + \beta_{353} CpG_{353}$+error

The coefficient values can be found in Horvath's Additional file 3. Based, on the coefficient values from the regression model, DNAmAge was estimated as follows

$$DNAmAge = inverse.F(\beta_0 + \beta_1 CpG_1 + \cdots + \beta_{353} CpG_{353})$$

Thus, the regression model can be used to predict the transformed age value by simply plugging the beta values of the selected CpGs into the formula. The linear part, (i.e. the weighted average of the selected CpGs) is visualized as red line in Figure 3.2. As expected, the red line passes through the weighted average of the CpGs since the weighted average of the methylations should be correlated with age.

Figure 3.2 The weighted average of the 353 clock CpGs versus chronological age in the training data sets.



The "*glmnet"* function required the user to specify two parameters (alpha and beta). The elastic-net regression model required alpha to be set to 0.5. However, the lambda value of

0.02255706 was chosen by applying a 10 fold cross validation to the training data (via the R function cv.glmnet).

The following R code provides details on the analysis.

```
library(glmnet)
# use 10 fold cross validation to estimate the lambda parameter
# in the training data
glmnet.Training.CV = cv.glmnet(datMethTraining, F(Age),
nfolds=10,alpha=alpha,family="gaussian")
# The definition of the lambda parameter:
lambda.glmnet.Training = glmnet.Training.CV$lambda.min
# Fit the elastic net predictor to the training data
glmnet.Training = glmnet(datMethTraining, F(Age), family="gaussian", alpha=0.5,
nlambda=100)
# Arrive at an estimate of of DNAmAge
DNAmAgeBasedOnTraining=inverse.F(predict(glmnet.Training,datMeth,type="response",s=la
mbda.glmnet.Training))
```

3.3 Predictive accuracy measures

In order to validate DNAm age (predicted biological age), Horvath considered several measures of predictive accuracy. The first, referred to as ‘age correlation’, was the Pearson correlation coefficient between DNAm age and chronological age where higher values correspond to greater predictive accuracy. Yet, it had the following limitations: it cannot be used for studying whether DNAm is well calibrated, it cannot be calculated in data sets whose subjects had the same chronological age (for example, cord blood samples from newborns), and it was strongly dependent on the standard deviation of age (as described below). The second accuracy measure, referred to as (median) ‘error’, was the median of the absolute difference between DNAm age and chronological age. Thus, a test set error of 3.6 years indicated that DNAm age differs by less than 3.6 years in 50% of subjects. The error was well suited for studying whether DNAm age is poorly calibrated. Lastly, average age acceleration defined as the average difference between DNAm age and chronological age, can be used to determine whether the DNAm age of a given tissue was consistently higher (or lower) than expected.

3.4 Chapter 3 References

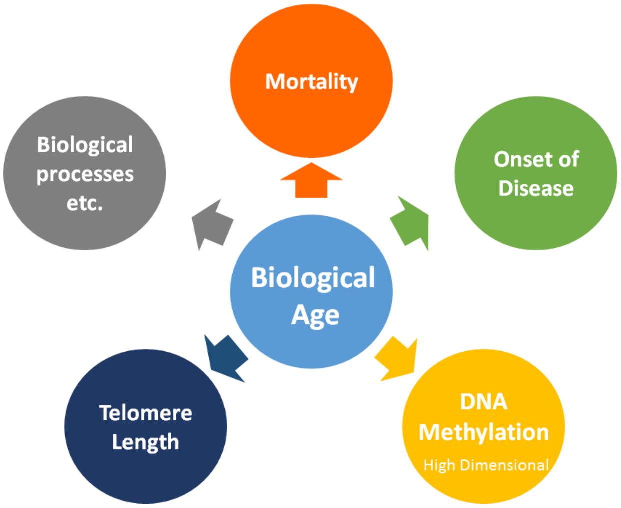
Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome biology*, *14*(10), R115.

Chapter 4: Ensemble Model Based on the Klemera and Doubal Approach (Ensemble KD)

4.1 Introduction

In estimating the latent trait (i.e. biological age), there are two areas of discussion, the estimating algorithm and the selection of the predictors/traits/biomarkers. The Klemera and Doubal (KD) approach has been shown to provide accurate estimates of biological age (Klemera and Doubal, 2009; Cho et al. 2010; Levine 2012) and, thus, is a reliable estimator of biological age. In selecting the biomarkers to estimate biological age, Cho et al. (2010) and Levine (2013) chose functionally independent biomarkers as suggested by Klemera and Doubal (2010). However, adaptations of the KD approaches are needed with the advancement of technology where functionally independent biomarkers may be harder to define such as in DNA methylation (Figure 4.1).

Figure 4.1: New methods are needed to estimate biological age with advancement in technology.



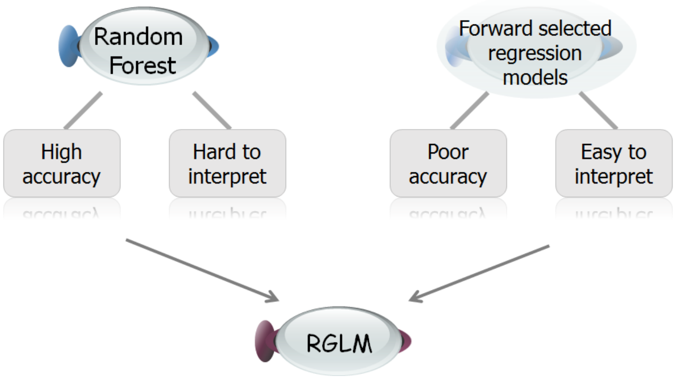
Horvath et al. (2012) had successfully identified DNA methylation co-modules related to aging. Hence, DNA methylation data can be used in aging research. However, DNA methylation data are high dimensional data expressing many complex systems that are currently unknown in literature. Consequently, finding "functionally independent" CpG's are not simple and new methods are needed to utilize DNA methylation data in estimating biological age.

In genetic studies, it is often difficult to pre-identify genes of interest without encountering multiple testing issues and evaluating all the genes is also not a practical approach. When evaluating the characteristics of the KD approaches, we found that when there are few predictors or when the data signal is weak, the estimates from the KD approach can be unstable and leads to high variability. Hence, we explored whether an ensemble model based on the KD approach is more robust and more accurate to estimating biological age.

4.2 Background Method

Our ensemble model based on the KD approach mirrors the random GLM method proposed by Song, Langfelder, & Horvath (2013). The random GLM encompasses the ideology of random forest and generalized linear models. In short, random forest is an ensemble model by combining a multitude of tree models such that each tree depends on the values of a random vector that are sampled independently and with the same distribution for all trees in the forest (Breiman, 2001). Ensemble models such as random forest are known for its highly predictive accuracy (Breiman, 1996; Breiman, 2001). Though random forests have superior accuracy, random forest models are often hard to interpret since the associations between the biomarkers/predictors and the outcome are not always transparent. On the other hand, generalized linear models are highly interpretable since generalized linear models provide estimated coefficients that guide researchers to understand the relationships between the predictor(s) and the outcome of interest. They also incorporate easy to understand variable selection algorithms such as forward selection. Yet, the ease of interpretability for forward selected regression models comes with a cost. Forward variable selection and other straightforward variable selection methods easily overfit the data which results in unstable and inaccurate predictions. By combining characteristics from ensemble models with generalized linear model, Song et al. (2013) achieved a highly accurate and interpretable model, random generalized linear models (RGLM) (Figure 4.2) which we used as the guide for our ensemble KD approach.

Figure 4.2: Rational behind RGLM, combining the logic of random forest with generalized linear model with forward select variable selection.



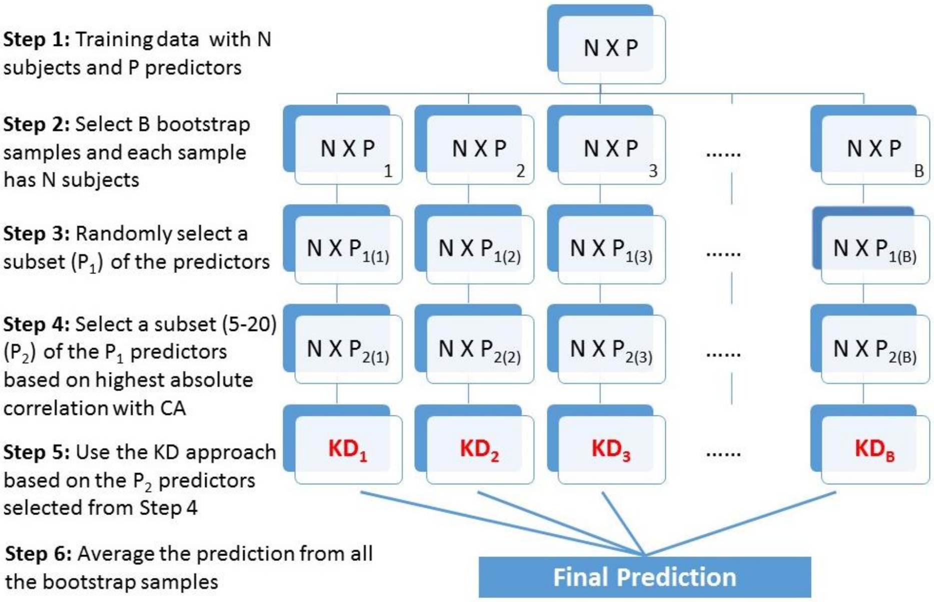
4.3 Ensemble KD Model with Bootstrap Algorithm

The flowchart of the ensemble KD model is presented in Figure 4.3 which is a mirror image of the random GLM overview. One method for constructing an ensemble predictor is using bootstrap aggregation (bagging). Our new ensemble model incorporates bootstrapping in order to arrive at more stable estimates by creating multiple versions of the data generated through bootstrapping from the original data, and the observations are randomly sampled with replacement.

Within each bootstrap sample, an individual KD model is constructed along with the predictions. The final prediction is computed by averaging the predictions across all bootstrap samples.

4.3.1 Overview of Ensemble KD Approach

Figure 4.3: Flowchart of the ensemble model based on the KD approach



Step 1: Training data with N subjects and P predictors
Step 2: Select B bootstrap samples and each sample has N subjects
Step 3: Randomly select a subset (P_1) of the predictors
Step 4: Select a subset (5-20) (P_2) of the P_1 predictors based on highest absolute correlation with CA
Step 5: Use the KD approach based on the P_2 predictors selected from Step 4
Step 6: Average the prediction from all the bootstrap samples
N X P
N X P_1
N X P_2
N X P_3
......
N X P_B
N X $P_{1(1)}$
N X $P_{1(2)}$
N X $P_{1(3)}$
N X $P_{1(B)}$
N X $P_{2(1)}$
N X $P_{2(2)}$
N X $P_{2(3)}$
N X $P_{2(B)}$
KD_1
KD_2
KD_3
KD_B
Final Prediction

The “*steps*” described in this section correspond to the steps in the flowchart (Figure 4.3). The KD approach generates 3 types of prediction for biological age: average of the traits/predictors, KD approach without chronological age as one of the predictors (KD1), and KD approach with chronological age as one of the predictors (KD2). Hence, the matrices “datyTRUE1Boot”, “datyTRUE3Boot”, and “datyTRUE3Boot” store the 3 types of predictions from each bootstrap sample (i.e. bag). Consequently, the matrices will need to have the same number of rows as the number of the observations in the training dataset and the same number of columns as the number of bags (*n.boot*). In addition, “SD.yTRUE2Boot” and “SD.yTRUE3Boot” are two vectors that will store the standard deviation for the predictions based on KD1 and KD2 for each bag. Consequently, the two vectors have the length size equal to the number of bags. Additional by-products from the bootstrap samples such as individual correlations between the predictions and the predictors are also stored. In addition, empty matrices are pre-defined in order to store the bootstrap samples and the randomly selected predictors for each bag. The following R code, in gray font, provides details on the analysis.

```
datyTRUE1Boot=matrix(NA,nrow=n,ncol=n.boot)
datyTRUE2Boot=matrix(NA,nrow=n,ncol=n.boot)
datyTRUE3Boot=matrix(NA,nrow=n,ncol=n.boot)

SD.yTRUE2Boot=matrix(NA,nrow=n.boot,ncol=1)
SD.yTRUE3Boot=matrix(NA,nrow=n.boot,ncol=1)

p.max1=rep(NA,n.boot) ; datCorMax1=matrix(NA,nrow=n.boot,ncol=n.trait)
p.max2=rep(NA,n.boot); datCorMax2=matrix(NA,nrow=n.boot,ncol=n.trait)
p.max3=rep(NA,n.boot); datCorMax3=matrix(NA,nrow=n.boot,ncol=n.trait)

datCorMean1=matrix(NA,ncol=n.trait,nrow=n.boot)
datCorMean2=matrix(NA,ncol=n.trait,nrow=n.boot)
datCorMean3=matrix(NA,ncol=n.trait,nrow=n.boot)
```

In *step 1*, the training data is assumed to have *N* observations with *P* predictors. There are four input parameters that need to be user specified: the number of bootstrap samples (*n.boot*), the number of predictors selected (*n.covariates*).

```
# R code for inputting parameters
n.boot=500
n.covariates=round(0.2*n.trait,0).
```

In *step 2*, the current default for the input parameters are 500 bootstrap samples.

```
# Start of Bootstrap #
for (i in 1:n.boot) {

set.seed(i)
indexBoot=sample(1:n,n, replace=TRUE)
#select the bootstrap samples
```

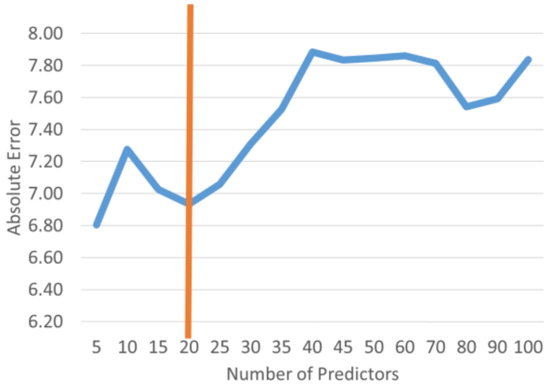
```
set.seed(i)
n.topCovariate=round(ifelse(n.covariates<=20,runif(1,2,5),runif(1,5,20)),0)
```

Within each bag, we sample with replacement the same number of observations (size *N*) as the training data. Since each bootstrap sample is generated at random, a random seed is set at each random sampling in order to produce consistent results for each analysis and, at the same time, different across bags. At *step 3*, a random subset P_1 of the *P* predictors. Currently, the default for P_1 is set as 20% of the number of *P* predictors are selected at *step 2*.

```
set.seed(i)
indexCovBoot=sample(1:n.trait,n.covariates, replace=FALSE)
#select the bootstrap covariates
indexOOB=setdiff(1:n, indexBoot)
#select the out of bag samples
```

Next, another random subset of 5 to 20 of the selected P_1 predictors are selected as the top predictors in step 4. If 20% of the number of predictors is less than 20, then only 1 to 5 of the top predictors will be selected in *step 4*. We decide to randomly select 5 to 20 in step 4 because, in a separate simulation analysis, we try to determine the number of predictors to include in each of the KD models would be optimal. We find that including too many predictors in the KD models lead to higher error (Figure 4.4) and the number of predictors should be limited in each of the KD models. When the number of predictors ranges from 5 to 20, the absolute error is not stable and fluctuates. Hence, in the ensemble KD model, we allowed the model to randomly choose a number between 5 to 20 top predictors in each of the KD model.

Figure 4.4: Evaluating number of predictors to include in KD models



Then a random sampling of the traits (without replacement) is selected and bi-weight mid-correlations (Wilcox 2005, Section 9.3.8, page 399) between the chosen predictors and chronological age are computed. We select the top 5 to 20 (pre-specified input parameter: *n.topCovariate*) predictors that are the most correlated with chronological age as the input predictors in each of the KD model.

```
datXBoot0=datX[sort(indexBoot),sort(indexCovBoot)]
yBoot=y[sort(indexBoot)]
#puts the bootstrap samples and bootstrap covariates into a matrix
covSelectCor=order(abs(bicor(datXBoot0,yBoot)))[(n.covariates-(n.topCovariate-
1)):n.covariates]

matchColNames=match(colnames(datXBoot0[,sort(covSelectCor)]),colnames(datX))

datXBoot=datX[sort(indexBoot),matchColNames]
#puts the bootstrap samples and bootstrap covariates into a matrix
datXOOB=datX[sort(indexOOB),matchColNames]
#select the out of bag samples and bootstrap covariates into another matrix

selectedSamples=unique(sort(indexBoot))
```

Within each bag, the KD approach uses the bootstrap sample and the top 5 to 20 most correlated traits (with chronological age) based on the random subset of the original traits. If the number of traits is less than 20, then only the top 1 to 5 most correlated traits will be selected in each bag.

The estimate biological age, standard deviations of the estimated biological age from the KD approach are stored in the previous created empty matrices in *step 5*.

```
TTBoot=TrueTrait(datX=datXBoot,y=yBoot,datXtest=datXOOB)
yTRUE1Boot=TTBoot$datEstimates[,2];
yTRUE1Boot=cbind(as.vector(sort(indexBoot)),yTRUE1Boot);
yTRUE1Boot=yTRUE1Boot[!duplicated(yTRUE1Boot[,1]),]; yTRUE1Boot=yTRUE1Boot[,-1]

yTRUE2Boot=TTBoot$datEstimates[,3];
yTRUE2Boot=cbind(as.vector(sort(indexBoot)),yTRUE2Boot);
yTRUE2Boot=yTRUE2Boot[!duplicated(yTRUE2Boot[,1]),]; yTRUE2Boot=yTRUE2Boot[,-1]

yTRUE3Boot=TTBoot$datEstimates[,4];
yTRUE3Boot=cbind(as.vector(sort(indexBoot)),yTRUE3Boot);
yTRUE3Boot=yTRUE3Boot[!duplicated(yTRUE3Boot[,1]),]; yTRUE3Boot=yTRUE3Boot[,-1]

datyTRUE1Boot[selectedSamples,i]=yTRUE1Boot
datyTRUE2Boot[selectedSamples,i]=yTRUE2Boot
datyTRUE3Boot[selectedSamples,i]=yTRUE3Boot

SD.yTRUE2=TTBoot$SD.ytrue2
SD.yTRUE3=TTBoot$SD.ytrue3

SD.yTRUE2Boot[i]=rbind(TTBoot$SD.ytrue2)
SD.yTRUE3Boot[i]=rbind(TTBoot$SD.ytrue3)
```

The correlations between each of the selected traits with the estimated biological trait in each bag are also stored in the "*p.max*" matrices for the variable importance plot.

```
# calculate the max correlation between simulate outcome and the covariates

p.max1[i]=max.col(abs(cor(datyTRUE1Boot[,i],datXBoot,use="p")))
datCorMax1[i,sort(matchColNames)[p.max1[i]]]=1
p.max2[i]=max.col(abs(cor(datyTRUE2Boot[,i],datXBoot,use="p")))
datCorMax2[i,sort(matchColNames)[p.max2[i]]]=1
p.max3[i]=max.col(abs(cor(datyTRUE3Boot[,i],datXBoot,use="p")))
datCorMax3[i,sort(matchColNames)[p.max3[i]]]=1

datCorMean1[i,sort(matchColNames)]=cor(datyTRUE1Boot[,i],datXBoot,use="p")
datCorMean2[i,sort(matchColNames)]=cor(datyTRUE2Boot[,i],datXBoot,use="p")
datCorMean3[i,sort(matchColNames)]=cor(datyTRUE3Boot[,i],datXBoot,use="p")
}
# end of bootstrap
```

The iteration repeats for *n.boot* times and the predictions from each bag are stored.

```
yTRUE1BootMean=NULL
yTRUE2BootMean=NULL
yTRUE3BootMean=NULL

for (j in 1:n)  {
yTRUE1mean=mean(datyTRUE1Boot[j,],na.rm=T)
yTRUE2mean=mean(datyTRUE2Boot[j,],na.rm=T)
yTRUE3mean=mean(datyTRUE3Boot[j,],na.rm=T)

yTRUE1BootMean = rbind(yTRUE1BootMean,c(yTRUE1mean))
yTRUE2BootMean = rbind(yTRUE2BootMean,c(yTRUE2mean))
yTRUE3BootMean = rbind(yTRUE3BootMean,c(yTRUE3mean))
}
datyTRUE1 = as.vector(yTRUE1BootMean);
datyTRUE2 = as.vector(yTRUE2BootMean);
datyTRUE3 = as.vector(yTRUE3BootMean);

datyTRUEBoot=cbind(y,yTRUE,datyTRUE1,datyTRUE2,datyTRUE3)
```

Finally, the predictions of each KD approach (one per bag) are averaged across bags to arrive at the final prediction and are stored in the matrix "*datyTRUEBoot*" in *step 6*. The predictions based on ensemble KD1 approach (without including chronological age as one of the predictors) are defined as "Boot.KD1" and, similarly, the predictions based on ensemble KD2 approach (including chronological age as one of the predictors) are defined as "Boot.KD2" in the later sections.

4.4 Variable Importance Plots

Additional by products such as variable importance measures are also constructed as part of the algorithm. Two types of variable importance measures are calculated, one based on the max correlation and the other based on the average correlation between the traits and the estimated biological age.

Within each bootstrap sample, the top candidate predictors selected as the input predictors for the KD approach are correlated with the estimated biological age using the bi-weight midcorrelation and the correlations stored in the "datCorMean1," "datCorMean2," and "datCorMean3" matrices. The predictor that has the highest absolute correlation with the predicted biological age within each bag is given one vote. This voting process is repeated throughout all bootstrap samples and the votes are stored in the "datCorMax1," "datCorMax2," and "datCorMax3" matrices across all bootstrap samples. At the end, all the traits will be ranked based on the number of votes (the number of times each predictor had the highest correlation with the predicted biological age) or the highest average correlation with the predicted biological ages across the bags. The resulting ranks will then be used to generate the variable importance plots to help determine which traits are most associated with the aging process.

```
p.mean1=NULL
p.mean2=NULL
p.mean3=NULL

for (k in 1:n.trait)  {
p1=mean(abs(datCorMean1[,k]),na.rm=T)
p2=mean(abs(datCorMean2[,k]),na.rm=T)
p3=mean(abs(datCorMean3[,k]),na.rm=T)

p.mean1=rbind(p.mean1,c(p1))
p.mean2=rbind(p.mean2,c(p2))
p.mean3=rbind(p.mean3,c(p3))
}

p.mean=cbind(p.mean1,p.mean2,p.mean3)

nameDatPMean=rep(NA,n.trait)
for (k in 1:n.trait){
nameDatPMean[k]=paste("Var",k)
}
rownames(p.mean)=nameDatPMean

sumCorMax1=rep(NA,n.trait)
sumCorMax2=rep(NA,n.trait)
```

```
sumCorMax3=rep(NA,n.trait)

# Plot of
for (k in 1:n.trait) {
sumCorMax1[k]=sum(datCorMax1[,k],na.rm=T)
sumCorMax2[k]=sum(datCorMax2[,k],na.rm=T)
sumCorMax3[k]=sum(datCorMax3[,k],na.rm=T)
}

sumCorMax1=matrix(sumCorMax1); rownames(sumCorMax1)=nameDatX
sumCorMax2=matrix(sumCorMax2); rownames(sumCorMax2)=nameDatX
sumCorMax3=matrix(sumCorMax3); rownames(sumCorMax3)=nameDatX

CorMax1=sumCorMax1[order(sumCorMax1[,1]),]
CorMax2=sumCorMax2[order(sumCorMax2[,1]),]
CorMax3=sumCorMax3[order(sumCorMax3[,1]),]
```

As mentioned previously, two types of variable importance plots are generated as by-products of the bootstrap algorithms: one based on the highest absolute correlation between the traits and the estimated biological age and the other based on the average correlation between the traits and the estimated biological age across all bags. The top 20 traits that have the highest number of votes and the highest average correlation across bags are presented in the plots. Figure 4.5 presents an illustrated example the discussed variable importance plots of a simulated data with 200 traits.

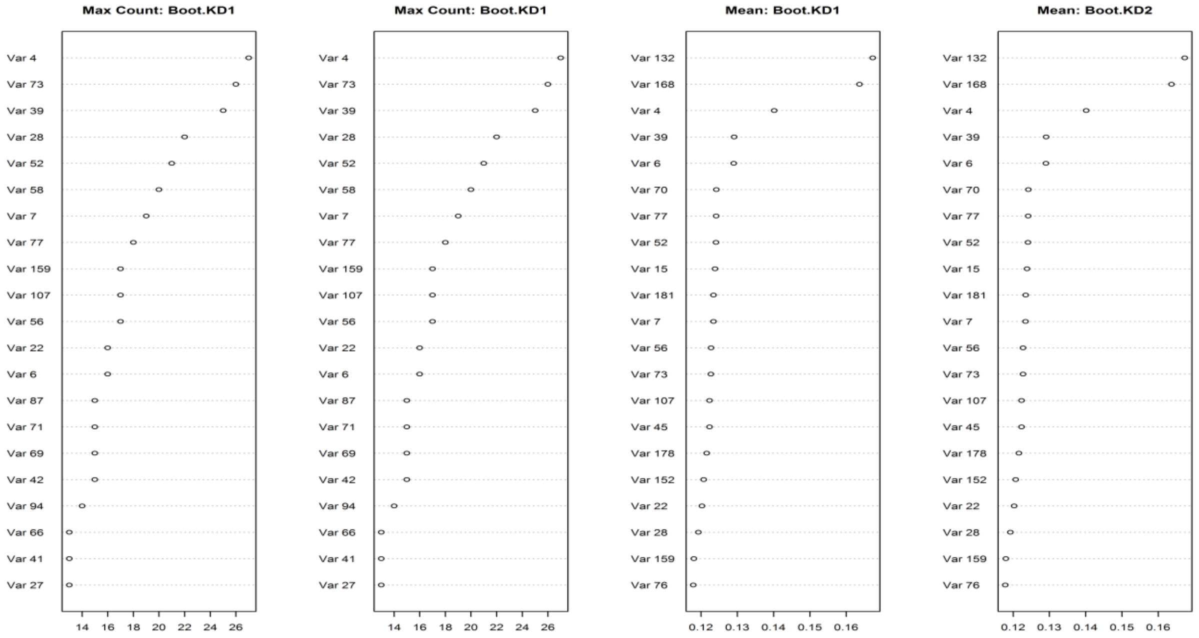
```
#pdf("VIP Corr.pdf")

par(mfrow=c(1,4))
#dotchart(CorMax1[(length(CorMax1)-
20):length(CorMax1)],labels=names(CorMax1[(length(CorMax1)-
20):length(CorMax1)]),main="Max Count: Sim. Truth 1")
dotchart(CorMax2[(length(CorMax2)-
20):length(CorMax2)],labels=names(CorMax2[(length(CorMax2)-
20):length(CorMax2)]),main="Max Count: Boot.KD1")
dotchart(CorMax3[(length(CorMax3)-
20):length(CorMax3)],labels=names(CorMax3[(length(CorMax3)-
20):length(CorMax3)]),main="Max Count: Boot.KD1")

#dotchart(sort(p.mean[,1])[(length(sort(p.mean[,1]))-
20):length(sort(p.mean[,1]))],labels=names(sort(p.mean[,1])[(length(sort(p.mean[,1]))-
20):length(sort(p.mean[,1]))]),main="Mean Corr: Sim. Truth 1")
```

```
dotchart(sort(p.mean[,2])[(length(sort(p.mean[,2]))-
20):length(sort(p.mean[,2]))],labels=names(sort(p.mean[,2])[(length(sort(p.mean[,2]))-
20):length(sort(p.mean[,2]))]),main="Mean: Boot.KD1")
dotchart(sort(p.mean[,3])[(length(sort(p.mean[,3]))-
20):length(sort(p.mean[,3]))],labels=names(sort(p.mean[,3])[(length(sort(p.mean[,3]))-
20):length(sort(p.mean[,3]))]),main="Mean: Boot.KD2")
dev.off()
```

Figure 4.5 Variable importance plot based on the average or max correlation between individual traits with predicted outcome.



4.5 Predictive Accuracy Measures

In order to evaluate the performance of the ensemble KD models, the accuracy measurements used will be similar to the accuracy and error measures described in Horvath (2013) with minor adjustments. To circumnavigate the dependency on the standard deviation of age, Spearman correlation is used as an estimate of accuracy rather than Pearson correlation since the Spearman correlation utilizes the ranks of the measure and is not dependent on the actual standard deviation of age. In addition, the median absolute difference between estimated biological trait and chronological age or simulated true biological age if available will also be used as an estimate of error.

4.6 Chapter 4 References

Breiman, L. (1996). Bagging predictors. *Machine learning*, *24*(2), 123-140.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Cho, I. H., Park, K. S., & Lim, C. J. (2010). An empirical comparative study on biological age estimation algorithms with an application of Work Ability Index (WAI). *Mechanisms of ageing and development*, *131*(2), 69-78.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., ... & Ophoff, R. A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol*, *13*(10), R97.

Klemera, P., & Doubal, S. (2006). A new approach to the concept and computation of biological age. *Mechanisms of ageing and development*,*127*(3), 240-248.

Levine, M. E. (2012). Modeling the rate of senescence: can estimated biological age predict mortality more accurately than chronological age?. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *68*(6), 667-674.

Song L, Langfelder P, Horvath S (2013) Random generalized linear model: a highly accurate and interpretable ensemble predictor. BMC Bioinformatics 14:5 PMID: 23323760 DOI: 10.1186/1471-2105-14-5.

Wilcox RR (2005). Introduction to Robust Estimation and Hypothesis Testing. 2nd edition. Academic Press.

Chapter 5: Simulation

5.1 Introduction

In this chapter, we apply the KD approach and the Robust KD approach described in Chapter 2 and the ensemble KD approach and the ensemble Robust KD approach described in Chapter 4 to predict biological age in a series of simulations in order to evaluate the accuracy and precision of the ensemble KD approaches. Other popular prediction models are also evaluated for all the simulated scenarios. The models evaluated are penalized (elastic-net) regression, principal component analysis, and random GLM. Other estimating models such as linear model (by itself), factor analysis with linear model, KD1, and ensemble KD1 are considered and evaluated. However, the results from those models just listed are not included in this section. Linear model is not suitable for the application we are interested in because of the large amount of predictors resulting in overfit issues and the estimates from factor analysis with linear model are very similar to principal component analysis with linear model. Hence, results for the linear model and factor analysis with linear model are not presented in this chapter. In addition, we found that the estimates from KD1 and KD2 are very similar. As a result, only the KD2 and the ensemble KD2 estimates are presented (these are referred throughout the text as KD approach and the ensemble KD approach).

The estimated biological age is evaluated against the true biological age and chronological age. Accuracy is measured by correlating the predicted biological age and the true biological age using Spearman correlation and the median absolute error as described in Chapter 4.5. Higher values of correlation and lower values of the median absolute error correspond to better accuracy.

5.2 Methods

Let CA be the observed chronological age and BA be the true biological age that is unobserved. CA is assumed to follow a normal distribution with mean 50 and standard deviation of 20.

$$CA \sim N(50,20)$$

```
y=rnorm(150,mean=50,sd=20) # y is CA
```

In addition, the true biological age follows the same assumption as Klemera and Doubal (2006) such that biological age is equivalent to the chronological age with some additional random variation with mean zero and variance s_B^2.

$BA = CA + R_B(0, s_B^2)$

where $R_B(0, s_B^2)$ is a random variable that follows a normal distribution with zero mean and variance s_B^2. In the simulations, we assum s_B^2 to be 100.

```
yTRUE =y +rnorm(150,sd=10) #yTRUE is true biological age
```

For every simulation, the sample size is fixed at 150. There are two main overarching scenarios, weak signal and strong signal. Within each scenario, we generate increasing number of traits ("simulated biomarkers") so that we can evaluate the trend of the predicted biological with increasing number of predictors. The simulated traits are simulated using the "*simulateModule*" function under the "WGCNA" package in R Cran. The weak signal scenario has covariates that are weakly correlated (r range: 0.01-0.2) with the true biological age, and similarly, the strong signal scenario has covariates that are strongly correlated (r range: 0.01-0.5) with the true biological age

```
# Weak Scenario
for (i in seq(20,190,10)){
nGenes=i
datX=simulateModule(yTRUE,nGenes=nGenes,
minCor=0.01,maxCor=0.2,geneMeans=rnorm(nGenes,50,30)  )
write.table(datX,paste("WeakDatX",nGenes,".csv",sep=""),sep=",",row.names=FALSE)
}
for (i in seq(200,1000,100)){
nGenes=i
datX=simulateModule(yTRUE,nGenes=nGenes,
minCor=0.01,maxCor=0.2,geneMeans=rnorm(nGenes,50,30)  )
write.table(datX,paste("WeakDatX",nGenes,".csv",sep=""),sep=",",row.names=FALSE)
}
# Strong Scenario
for (i in seq(20,190,10)){
nGenes=i
datX=simulateModule(yTRUE,nGenes=nGenes,
minCor=0.01,maxCor=0.5,geneMeans=rnorm(nGenes,50,30)  )
write.table(datX,paste(“ModerateDatX",nGenes,".csv",sep=""),sep=",",row.names=FALSE)
}

for (i in seq(200,1000,100)){
nGenes=i
datX=simulateModule(yTRUE,nGenes=nGenes,
minCor=0.01,maxCor=0.5,geneMeans=rnorm(nGenes,50,30)  )
write.table(datX,paste("ModerateDatX",nGenes,".csv",sep=""),sep=",",row.names=FALSE)
}
```

The estimated biological age is extracted from each prediction model and correlates with the true biological age and chronological age. In addition, the median of the absolute error between the estimated biological age with the true biological age or the chronological age is also calculated.

```
datX=read.csv(paste("C:/Users/Wendy/Desktop/Dissertation/Defense/Simulation/Moderate
Simulation Data/ModerateDatX",i,".csv",sep=""))
KDEstimate=TrueTrait(datX=datX,y=y)
KDPred=KDEstimate$datEstimates[,4]

RobustKDEstimate=TrueTraitRobust(datX=datX,y=y)
RobustKDPred=RobustKDEstimate$datEstimates[,4]
###
KD_corr=cor.test(KDPred, yTRUE,use="p",method="s")$estimate
KD_CAcorr=cor.test(KDPred, y,use="p",method="s")$estimate
KD_error=median(abs(KDPred-yTRUE),na.rm=T)
KD_CAerror=median(abs(KDPred-y),na.rm=T)
```

```
KD_var=var(KDPred,na.rm=T)

RobustKD_corr=cor.test(RobustKDPred, yTRUE,use="p",method="s")$estimate
RobustKD_CAcorr=cor.test(RobustKDPred, y,use="p",method="s")$estimate
RobustKD_error=median(abs(RobustKDPred-yTRUE),na.rm=T)
RobustKD_CAerror=median(abs(RobustKDPred-y),na.rm=T)
RobustKD_var=var(RobustKDPred,na.rm=T)
########################################################
# use 1st PC to estimate
pc=prcomp(datX, scale=TRUE)
pc.score=pc$x[,1] # 1st pc loadings
PCAPred=fitted(lm(y~pc.score))
PCAResidual=as.numeric(lm(PCAPred~y)$residual)

PCA_corr=cor.test(PCAPred, yTRUE,use="p",method="s")$estimate
PCA_CAcorr=cor.test(PCAPred, y,use="p",method="s")$estimate
PCA_error=median(abs(PCAPred-yTRUE),na.rm=T)
PCA_CAerror=median(abs(PCAPred-y),na.rm=T)
PCA_var=var(PCAPred,na.rm=T)
########################################################
alpha=0.5
cv.elasticfit = cv.glmnet(as.matrix(datX),y,nfolds=10,alpha=alpha,family="gaussian")
# The definition of the lambda parameter:
elastic.lambda = cv.elasticfit$lambda.min
elastic.fit = glmnet(as.matrix(datX),y, family="gaussian", alpha=alpha, nlambda=100)

# Arrive at an estimate of of DNAmAge
ElasticPred=predict(elastic.fit,as.matrix(datX),type="response",s=elastic.lambda)
ElasticResidual=as.numeric(lm(ElasticPred~y)$residual)

Elastic_corr=cor.test(ElasticPred, yTRUE,use="p",method="s")$estimate
Elastic_CAcorr=cor.test(ElasticPred, y,use="p",method="s")$estimate
Elastic_error=median(abs(ElasticPred-yTRUE),na.rm=T)
Elastic_CAerror=median(abs(ElasticPred-y),na.rm=T)
Elastic_var=var(ElasticPred,na.rm=T)
########################################################
RGLM = randomGLM(datX, y, nThreads = 1,keepModels=TRUE,classify=FALSE)
RGLMPred = RGLM$predictedOOB
RGLMResidual=as.numeric(lm(RGLMPred~y)$residual)

RGLM_corr=cor.test(RGLMPred, yTRUE,use="p",method="s")$estimate
RGLM_CAcorr=cor.test(RGLMPred, y,use="p",method="s")$estimate
RGLM_error=median(abs(RGLMPred-yTRUE),na.rm=T)
RGLM_CAerror=median(abs(RGLMPred-y),na.rm=T)
RGLM_var=var(RGLMPred,na.rm=T)
```

The simulation is repeated multiple times and each time with an increased number of traits (20 to 190 with increments of 10 and 200 to 1000 with increments of 100) in each scenario.

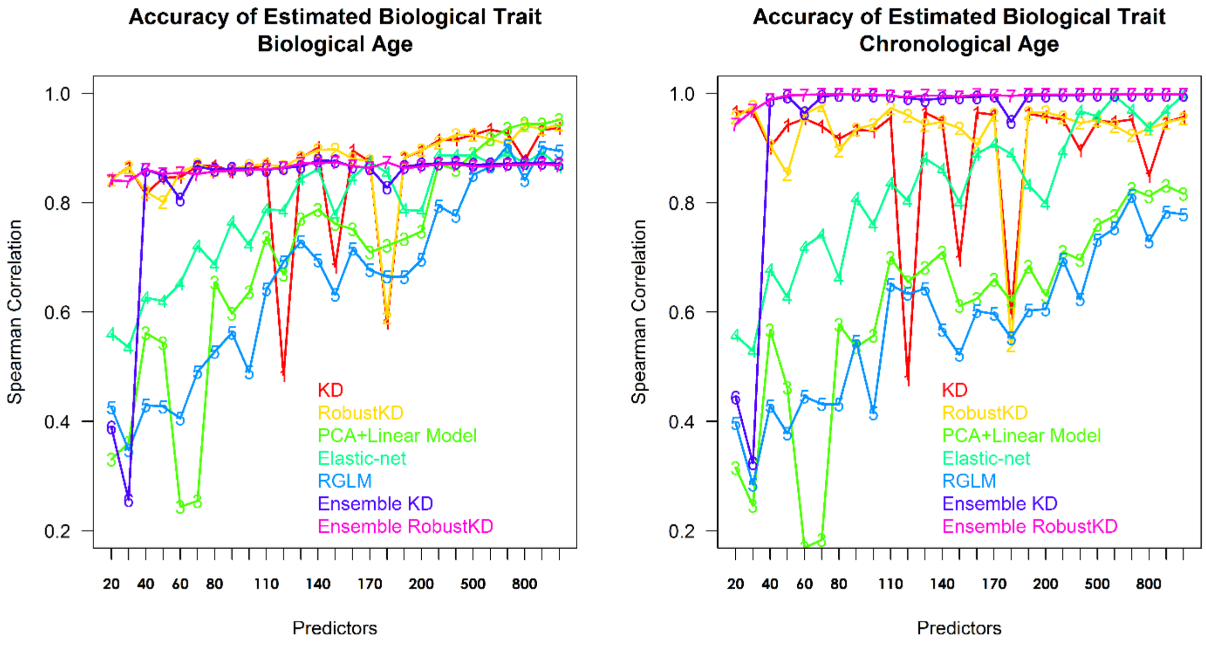
5.3 Results

The results of each main simulation (weak and strong) are presented in Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4 respectively. The left panel in each figure corresponds to the predictive accuracy measure between the predicted biological age and the true biological age. The right panel in each figure corresponds to the predictive accuracy measure between the predicted biological age and the chronological age. The x-axis for Figure 4.1, Figure 4.2, Figure 4.3 and Figure 4.4 corresponds to the number of predictors included in each prediction model. The y-axes for Figure 4.1 and Figure 4.3 correspond to the Spearman correlation between the estimated biological age and the true biological age or chronological age. The y-axes for Figure 4.2 and Figure 4.4 correspond to the median of the absolute difference between the estimated biological age and the true biological age or chronological age.

The predicted biological ages from KD approach and the ensemble KD approaches have superior correlation with both biological age and chronological age compared with principal component analysis with linear regression, elastic-net regression, and random GLM. However, as we mentioned before, the estimates of the KD approach can be unstable in the presence of outliers as we see in Figure 4.1 at 120, 160, and 180 traits where the correlation between the predicted biological age and the true biological age declined. Though using robust KD approach mitigated the imprecise estimates for two of those three scenarios, it is not without limitations. At the 180 weak predictors scenario, the robust KD approach was not able to produce robust estimates of biological age which lead to imprecise estimates of biological age. Further investigation reveal that the robust regression estimates in the reverse regression step are not stable since the estimates did not converge. However, the ensemble KD approach estimated biological age is stable throughout all simulations.

Results of the simulation indicate that increasing the number of predictors improved the accuracy (higher correlation between predicted biological age and true biological) for all prediction models and in both scenarios (weak and strong). In the weak scenario, the correlations start to plateau early for the KD approaches and the ensemble KD approaches, reaching high accuracy with fewer traits. Furthermore, increases in the number of traits also decrease the prediction error (reduction in the median absolute error for the estimated biological age. When the signal in the data is strong, all models have highly accurate predictions (high correlation and low error). In both the weak and strong scenario, the ensemble KD approaches has the highest correlation with chronological age and the lowest error compared to the other models.

Figure 4.1: Correlation between estimated biological ages with true biological age and chronological age in weak signal scenario results with varying number of covariates and models



Accuracy of Estimated Biological Trait
Biological Age

Accuracy of Estimated Biological Trait
Chronological Age

Figure 4.2: Median absolute error between estimated biological ages with true biological age and chronological age in weak signal scenario results with varying number of covariates and models

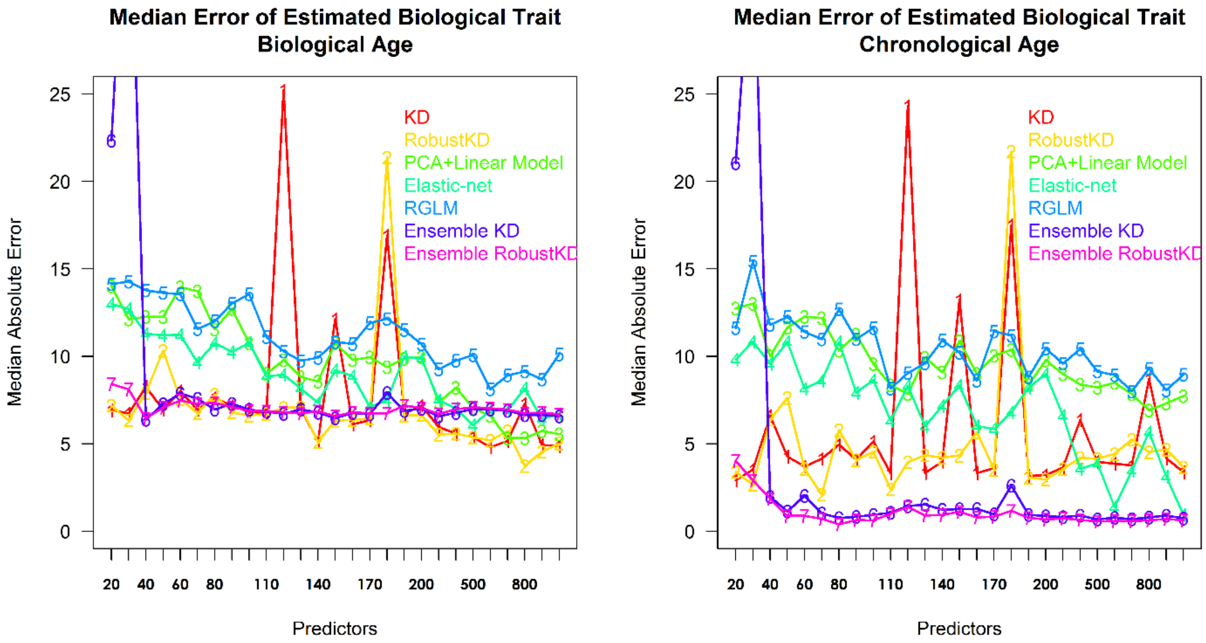


Figure 4.3: Correlation between estimated biological ages with true biological age and chronological age in strong signal scenario results with varying number of covariates and models

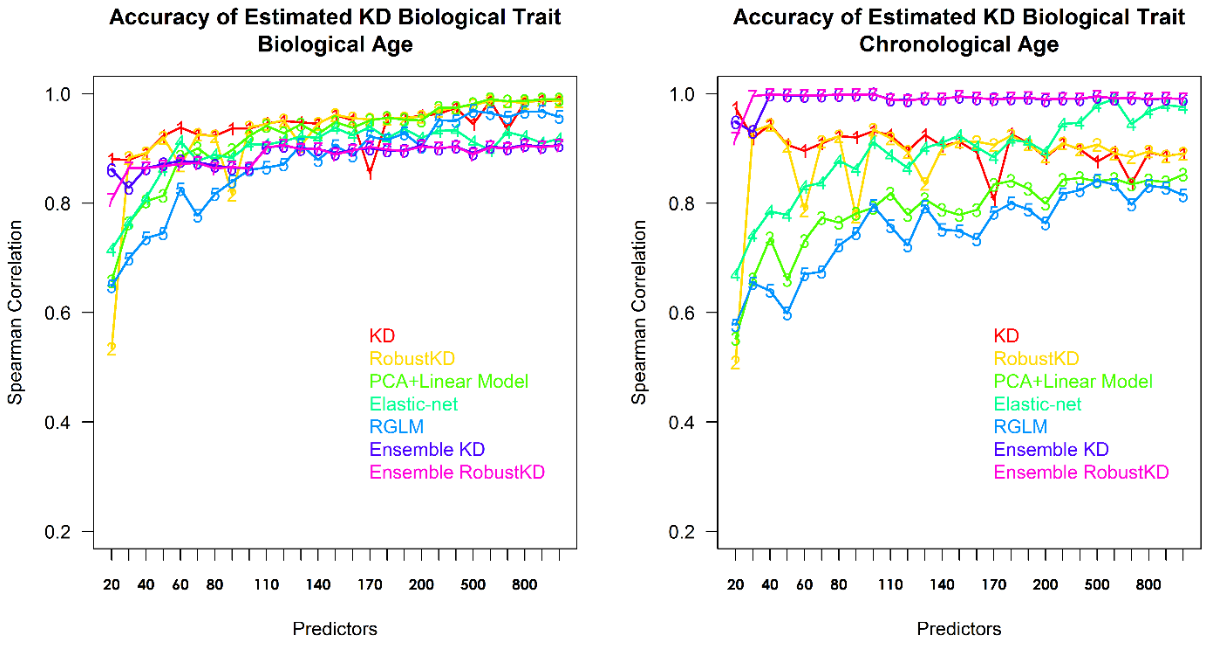
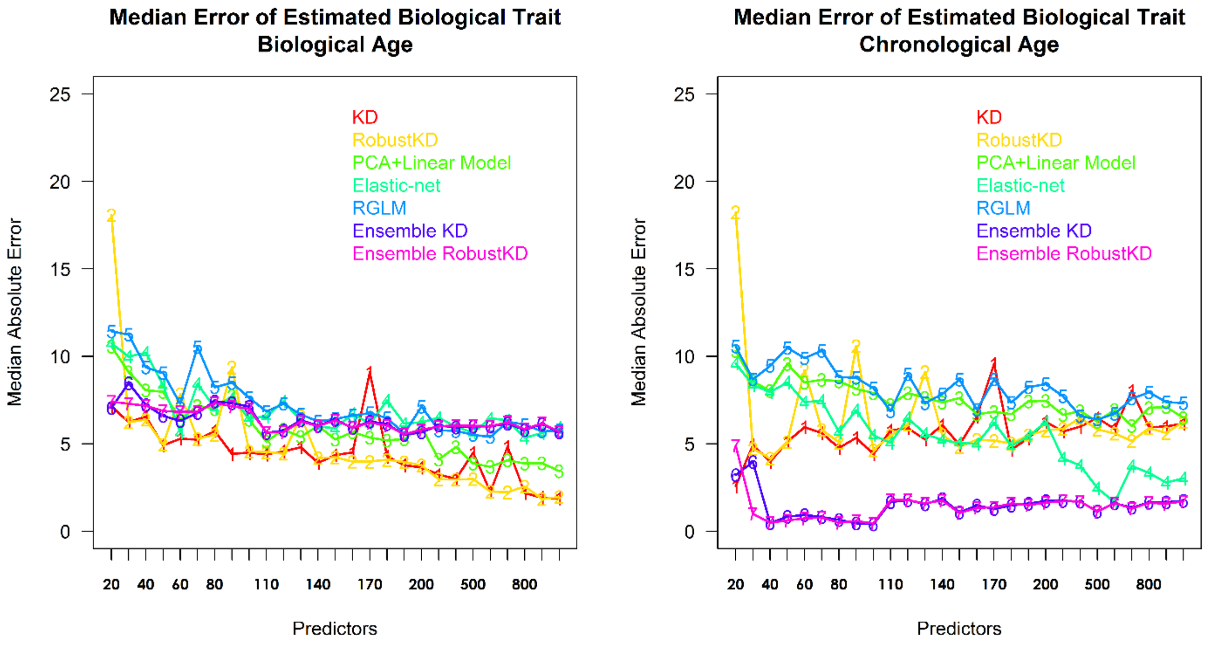


Figure 4.4: Median absolute error between estimated biological ages with true biological age and chronological age in strong signal scenario results with varying number of covariates and models



5.4 Conclusion

When the signal in the data is weak, the KD and the ensemble KD approaches both performed optimally well. They both have the highest accuracy (highest correlation and lowest prediction error) compared to all other methods. Though the ensemble KD approaches had slightly lower correlation with true biological age compared to KD approach, the difference is minimal (average difference in correlations was 0.028 in the weak scenario and 0.053 in the strong scenario). In summary, with increasing number of predictors, the KD approach had better prediction accuracy compared to existing methods (e.g. principal component analysis with linear model, elastic-net regression, random GLM). The ensemble KD approach provides more precise (less variable) but slightly less accurate (lower correlation with true biological age; higher median error) estimates of biological age compared to other KD approach.

5.5 Chapter 5 References

Klemera, P., & Doubal, S. (2006). A new approach to the concept and computation of biological age. *Mechanisms of ageing and development*,*127*(3), 240-248.

Chapter 6: Weighted Gene Correlation Network Analysis (WGCNA)

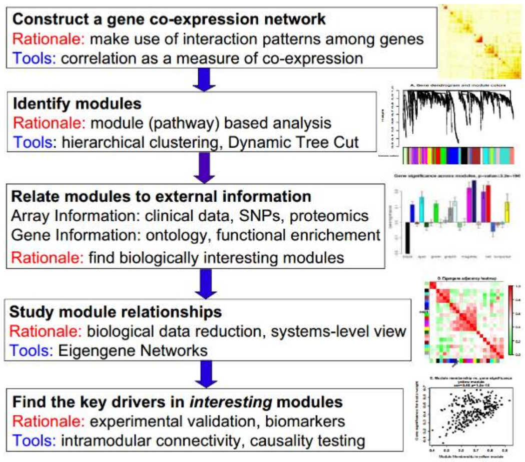
6.1 Introduction

Weighted gene co-expression network analysis (WGCNA) has been utilized in numerous studies ever since it was introduced in a comprehensive R package in 2008 (Langfelder & Horvath, 2008). WGCNA can be used in many ways such as a data exploratory tool or as a gene screening method. In short, WGCNA is a systems biology method for describing the correlation patterns among genes across microarray samples through network construction, module detection, gene selection, calculations of topological properties, data simulation, visualization, and interfacing with external software. By finding clusters (modules) of highly correlated genes, summarizing such clusters using the module eigengene or an intramodular hub gene, relating modules to one another and to external sample traits (using eigengene network methodology), and calculating module membership measures, researchers can generate testable hypothesis for validation in independent data sets.

6.2 Background

For the sake of simplicity, we will describe WGCNA in the context of gene expressions such as in microarray analyses, but WGCNA is not limited to only gene expression analyses. It can be applied to any high dimensional datasets. The steps of the network analysis will follow the same format as the process described in (Zhang & Horvath, 2005) and are depicted in Figure 6.1. In the application described in Chapter 4, we only utilized WGCNA to identify a disease associated module. Hence, this chapter will be dedicated to describing the preliminary steps of WGCNA that are pertinent to our application: constructing a co-expression network, identifying modules, and relating modules to external information. Detailed information and tutorials can be found at the WGCNA website.

Figure 6.1: Flow chart depicting WGCNA


Construct a gene co-expression network
Rationale: make use of interaction patterns among genes
Tools: correlation as a measure of co-expression
Identify modules
Rationale: module (pathway) based analysis
Tools: hierarchical clustering, Dynamic Tree Cut
Relate modules to external information
Array Information: clinical data, SNPs, proteomics
Gene Information: ontology, functional enrichement
Rationale: find biologically interesting modules
Study module relationships
Rationale: biological data reduction, systems-level view
Tools: Eigengene Networks
Find the key drivers in *interesting* modules
Rationale: experimental validation, biomarkers
Tools: intramodular connectivity, causality testing

6.2.1 Construct a gene co-expression network

Here we define a co-expression network as an undirected and weighted gene network where the nodes of the network may pertain to gene expression profiles or other continuous measurements. The edges are defined to be a measure of similarity or concordance between the gene expression profiles which will be denoted as s_{ij} where i and j are for gene i and gene j. As default, the edges are the absolute value of the Pearson correlation $s_{ij} = |cor(i,j)|$ and are calculated for all pair-wise comparisons of the gene expression across all samples. Researchers may consider other forms of correlations such as the bi-weight mid-correlation coefficient (Wilcox 2005, Section 9.3.8, page 399) instead of the Pearson correlation coefficient to protect against outliers. The only constraint for the similarity measure is that its values must lie between zero and one. The similarity measures are then stored in a similarity matrix designated by $S = [s_{ij}]$.

The correlation matrix is then transformed into a weighted undirected network by raising the absolute value of the pairwise correlations by a power of β. The function of raising the absolute value of the pairwise correlations by a power of β is known as an adjacency function, a_{ij}.

$$a_{ij} = power(s_{ij}, \beta) \equiv |s_{ij}|^{\beta}$$

There are many forms of adjacency functions since the adjacency function is simply a monotonically increasing function that maps the interval [0,1] to [0,1]. The adjacency function described here is a soft thresholding that preserves the continuous nature of the gene co-expression information, consequently, leading to more robust results and allowing for a simple geometric interpretation of network concepts.

In order to arrange the genes into clusters or modules, we used another measure of similarity known as the topological overlap of two nodes. The topological overlap matrix (TOM) $\Omega = [\omega_{ij}]$ is essentially a robust measure of interconnectedness.

$$\omega_{ij} = \frac{l_{ij} + a_{ij}}{min\{k_i, k_j\} + 1 - a_{ij}}$$

where $l_{ij} = \sum_u a_{iu} a_{uj}$, and $k_i = \sum_u a_{iu} = \sum_{j=1}^{n} a_{ij}$ is the node connectivity measure which is the sum of the connection strengths between a particular gene *i* and all other genes in the network. TOM can be easily transformed into a dissimilarity measure by subtracting each ω_{ij} value from one:

$$d_{ij}^{\omega} = 1 - \omega_{ij}.$$

6.2.2 Identify Gene Modules

WGCNA utilizes the TOM-based dissimilarity d_{ij}^{ω} to cluster the gene expression profiles by using average linkage hierarchical clustering. The cluster can be visualized using a hierarchical clustering tree (i.e. dendrogram) where the gene modules correspond to the branches.

6.2.3 Relate modules to external information

A key element in many network analyses is to relate the connectivity measure of the co-expression network to external information. Here we define a measure of module significance as the correlation between the external information and the module eigengene. Mathematically, the q^{th} module eigengene is the first principal component for the q^{th} module which can be considered as a composite measure for the module. Standard statistical methods such as regression models (e.g. linear regression) or multi-group comparison tests (e.g. Student t-test, ANOVA, etc.) can be used for evaluating module significance and is not limited to correlation analyses.

6.3 Chapter 6 References

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*, *9*(1), 559.

Wilcox RR (2005). Introduction to Robust Estimation and Hypothesis Testing. 2nd edition. Academic Press.

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, *4*(1).

Chapter 7: Accelerated Aging in Down Syndrome

7.1 Introduction

Down syndrome (DS) is a type of chromosomal disorder where an individual has a full or partial extra copy of chromosome 21. Each year, about 6,000 babies are born with DS in the United States, which is about 1 in every 700 babies born based on the latest release from the Center of Disease Control (Parker et al. 2010). DS continues to be the most common chromosomal disorder.

People with DS usually experience mild to moderate level of intellectual disability and are slower to speak compared to other children. In addition, adults with DS experience accelerated aging meaning they experience certain conditions and physical features that are common to typically aging adults at an earlier age than the general population such as premature skin wrinkling, greying of hair, hypogonadism, early menopause, hypothyroidism, declining immune function, and Alzheimer's disease. (Devenny et al., 2005; Patterson and Cabelof, 2012; Moran, 2013). With the advancement in current technology and the medical field, it is common for people with DS to reach old age, living well into their 50's, 60's and 70's. At the same time, longevity of life can also bring unforeseen challenges for adults with DS and their caregivers. Consequently, adults with DS, along with their families and caregivers, need accurate information and education about what to anticipate as a part of growing older, so that they may set the stage for successful aging (Moran, 2013).

However, biomarkers of aging are often limited resulting in difficulty to thoroughly evaluate whether DS is associated with accelerated aging effects. For example, telomere length is a popular candidate acting as a biomarker of aging. Vaziri et al. (1993) found evidence that linked decreased telomere length in DS subjects compared to control subjects and Jenkins et al. (2008)

suggested that decreased telomere length is associated with the presence of both dementia and mild cognitive impairment in adults with DS compared to those in age and sex matched controls with DS only. Nonetheless, one study found no association that cultured skin fibroblasts from DS subjects attained replicative senescence (a telomere length-dependent phenomenon) earlier than those from controls (Kimura et al., 2005). In addition, a review by Mather, Jorm, Parslow, and Christensen (2011), concluded that using telomere length as a biomarker of aging was inconclusive, telomere length was not a “universal” biomarker of aging, and hence, did not reflect general underlying aging processes. Therefore, using telomere length as a biomarker of aging in DS subjects may not be ideal.

In a recent study by Horvath et al. (2015), the authors utilized a quantitative molecular marker of aging known as the epigenetic clock to demonstrate that DS significantly increases the age of blood and brain tissue on average by 6.6 years. In addition, they observed significant age acceleration effects in brain (11 years) and blood (4 years) tissue. Our current study mimics the Horvath et al. (2015) study by utilizing a quantitative marker of aging estimated from the ensemble KD model to demonstrate that DS subjects experience an accelerated aging effect compared to the control subjects as a means to validate the performance of the ensemble KD model. However, our study is not meant to be compared with Horvath’s epigenetic clock since the epigenetic clocks uses specific 353 CpG methylation sites that may not be captured in the ensemble KD model. Consequently, the features use in the two models may not be the same and should not be directly compared. However, we plan to use the epigenetic clock as one of the comparison methods for evaluation purposes.

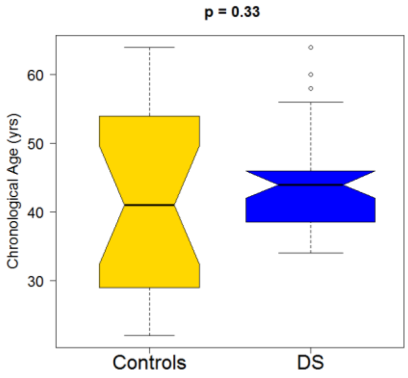
7.2 Objective

We aim to develop a new marker of aging as a means to measure disease progression. Since DS is a condition that occurs prior to birth, aside from chronological age, there are no measurements that measure the progression of the condition to the best of our knowledge. By applying the ensemble KD model, we anticipate to find that DS subjects will exhibit a highly significant age acceleration effect as is found by Horvath et al. (2015).

7.3 Data source and subjects

The DNA methylation is measured in DNA isolated from whole blood in Illumina 27K platform by Kerkel et al. (2010) (GSE25395). There is a total of 60 subjects. However, four subjects did not report their chronological age and were, thus, removed from the final sample resulting in a total of 56 subjects (35 DS and 21 controls) with an average age of 43 years old (range 22 to 64 years old). There is no significant difference in chronological age between the DS and control subjects (Figure 7.1: p=0.33). The x-axis is the DS status and the y-axis is the chronological age measured in years.

Figure 7.1: Distribution of chronological age by DS status.



7.4 Method

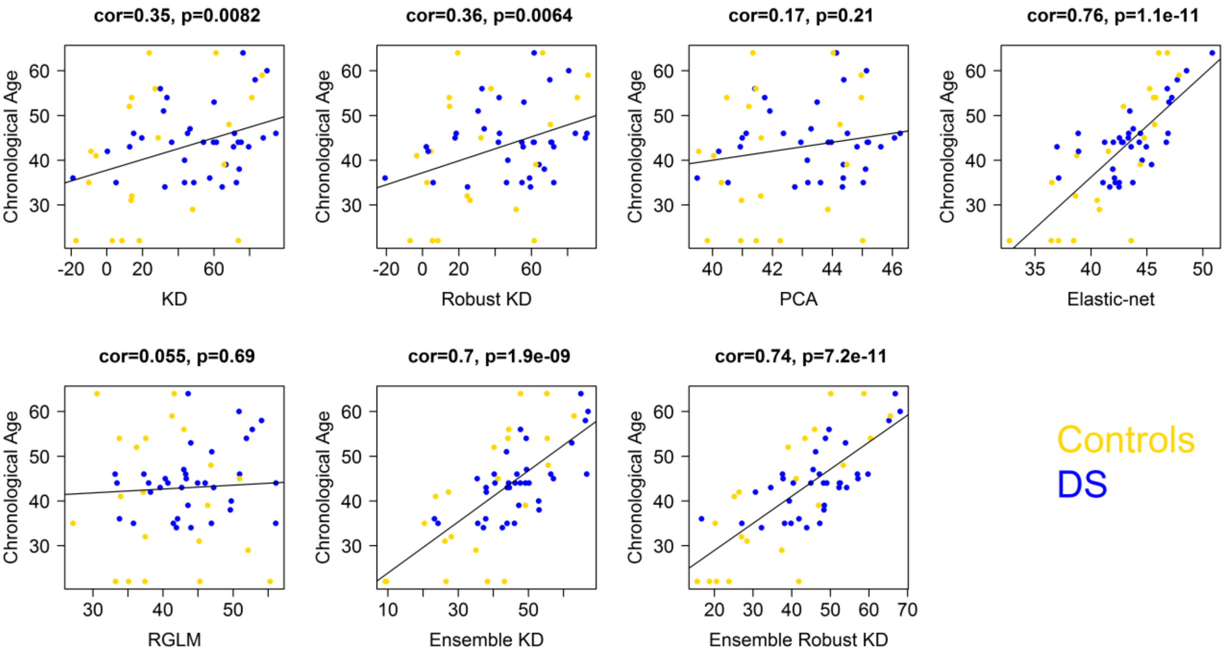
The DNA methylation levels are assessed in the peripheral blood leukocytes. Most of the CpG's have low variability and are not hyper/hypo methylated. Therefore, these CpG's are pre-screened out in order to save on computational time. Since we aimed to find a measurement of aging for the means to disease progression, we want to include CpG's that are related to disease in the ensemble KD model. Because we tried to estimate a latent variable based on hundreds of thousands of biomarkers, we wanted to pre-cluster these markers. Our hypothesis is that each latent variable give rise to a cluster of CpGs. To cluster the biomarkers, we use the benchmark method known as the weighted gene correlation network analysis (WGCNA). Other network analyses may also be used, but we choose WGCNA because it has been used successful at identifying co-methylation modules. (Song et al., 2012 and Horvath et al., 2012).

We use WGCNA to identify a co-methylation module that is strongly associated with DS status by evaluating the module eigengenes with condition status. Next we apply the ensemble KD models with the intramodule CpGs. The module identified includes a total of 226 CpG's and these CpG's are included in the ensemble KD models and other evaluation models as well (i.e. PCA with linear model, elastic-net regression, and RGLM). Lastly, we define measures of age acceleration as the residual resulting from a linear model that regressed the estimates from each of the models on chronological age since we expect the estimated values from the models (referred as predicted biological age) to be strongly correlated with chronological age (i.e. confounding). The KD, Robust KD, PCA with linear model and elastic-net models will be referred to as the simple models since the models do not utilize multiple iterations and the RGLM and the ensemble KD models will be referred to as the ensemble models in later text.

7.5 Results

The estimated values from each of the models are in all the models evaluated, the predicted biological age are unsurprisingly significantly correlated with chronological age (Figure 7.2). The x-axis is the estimated biological age based on the individual prediction model indicated in the x-axis label and the y-axis is the chronological age measured in years.

Figure 7.2: All the estimated biological traits are significantly correlated with chronological age. Hence, a chronological age adjusted of age accelerated needs to be used to compare between DS and control subjects.



Hence, we also use the residuals from linear model that regressed the estimated biologicals traits on chronological age to evaluate the age acceleration effect (i.e. adjusted aging acceleration). Figure 7.3 (non-ensemble models) and Figure 7.4 (ensemble models) shows that the estimated biological traits are higher among the DS subjects compared to the controls which provides evidence that DS subjects experience an accelerated aging effect compared to the controls in the single models. In the first row in Figure 7.3 and Figure 7.4, we compare the estimated biological age (y-axis) with disease status (x-axis) using Kruskal-Wallis test. In the second row of Figure 7.3 and Figure 7.4, we compare the adjusted aging acceleration (y-axis) with disease status (x-axis). The estimated age accelerations are significantly higher in the DS subjects compared to the controls in all models (p<0.05's) and was marginally significant in the Robust KD (p=0.06) and the elastic-net regression (p=0.089).

Figure 7.3: The estimated biological traits and the estimated age adjusted age acceleration are compared between DS and control subjects for all simple models. All models suggest an age accelerated effect for DS subjects.

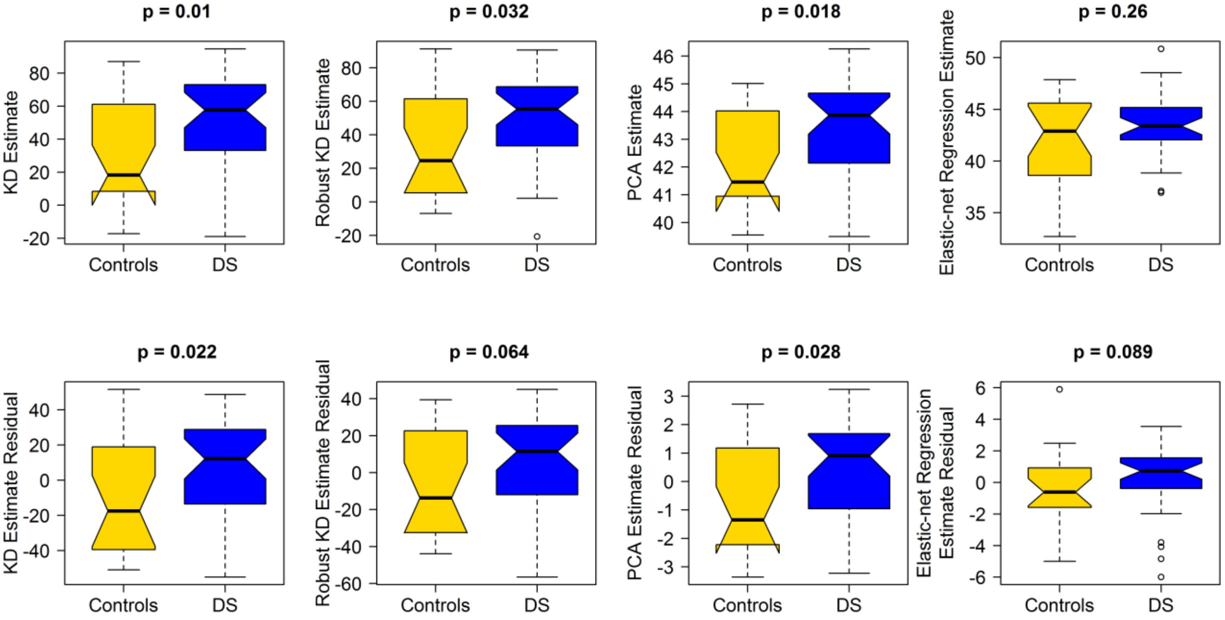
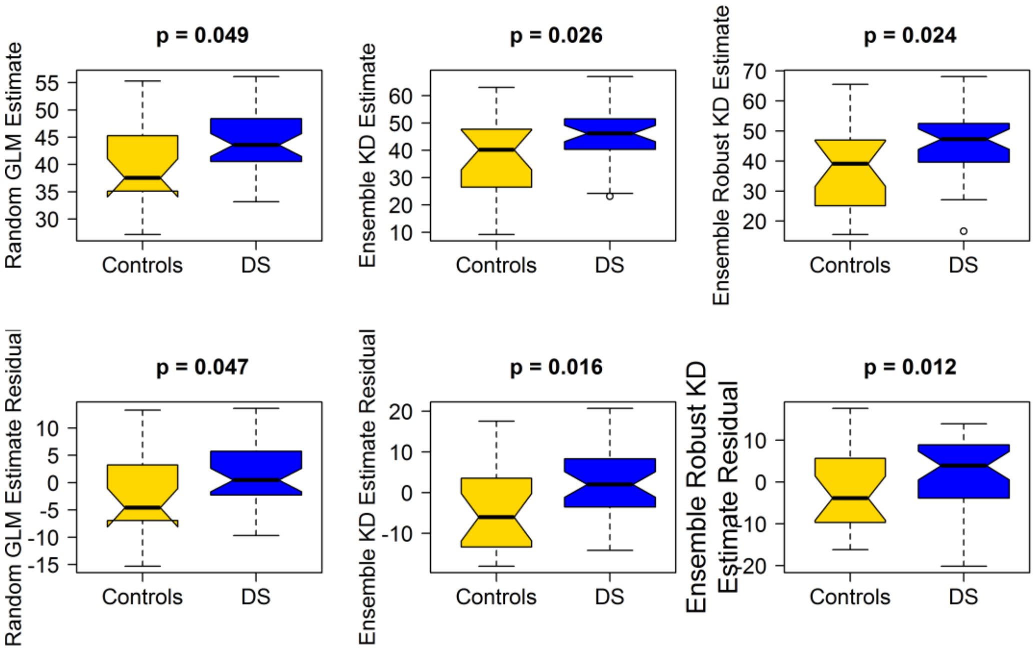


Figure 7.4 The estimated biological traits and the estimated age adjusted age acceleration are compared between DS and control subjects for all ensemble models. The ensemble models also suggest an age accelerated effect for DS subjects.



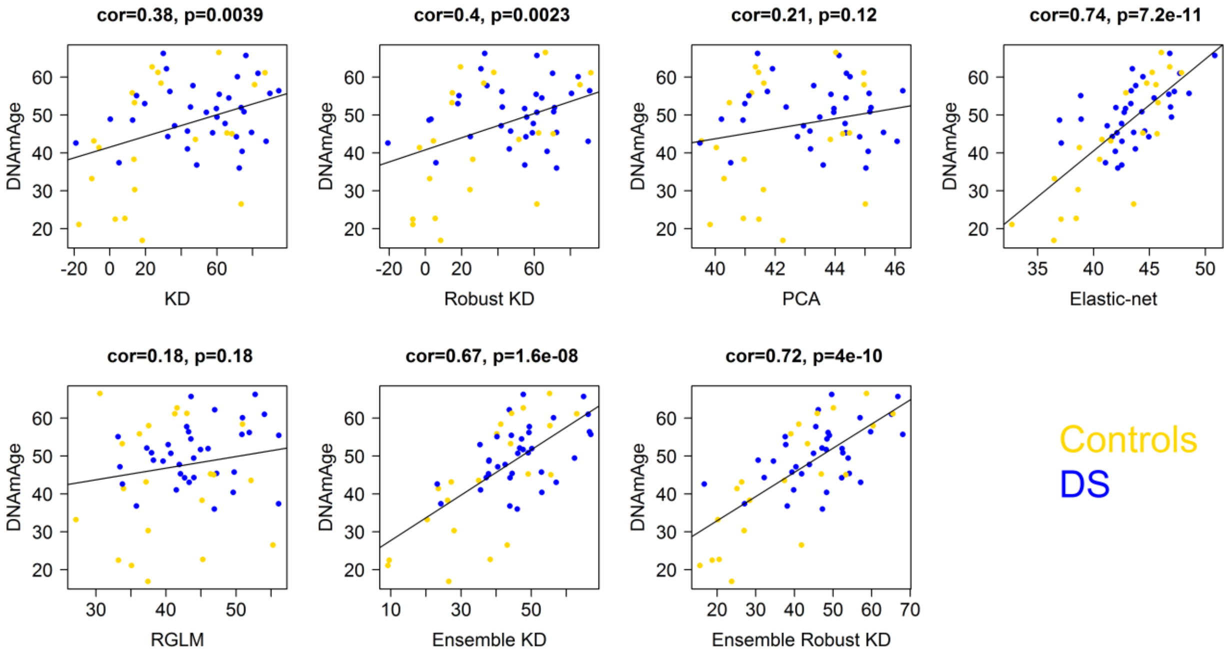
We evaluated the accuracy of our estimated biological trait with chronological age since a true biological age is not possible. The predictive accuracy measures for the estimated biological traits from each model are presented in Table 7.1. The ensemble KD models outperform the KD models in both accuracy measures (Spearman correlation and median error) by improving in the correlation by twice-fold and decreasing the median error by 70%. The PCA with linear model and the RGLM had poor accuracies though comparable median errors with the ensemble models. The elastic-net regression has the highest accuracy and lowest median error among all the models evaluated. Though we do not have a true biological age to accurately evaluate the predictability of our models, we do have the Horvath's DNAm age (Horvath, 2015) which has been shown to have superior accuracy in aging research. We correlate the predicted biological age with the DNAm age (Figure 7.5). The x-axis is the estimated biological age based on the individual prediction model indicated in the x-axis label and the y-axis is the DNAm age measured in years.

The results are similar to the findings presented in Table 7.1, the predictive accuracy measure with chronological age. The ensemble KD models had higher correlations with DNAm age than the KD models. The correlation between the elastic-net regression and the DNAm age is the highest among all the models evaluated.

Table 7.1: The predictive accuracy measures for the predicted biological age with chronological age. The ensemble KD models outperforms the KD models, PCA and RGLM.

Predictive Accuracy Measurements	**KD**	**Robust KD**	**PCA with Linear Model**	**Elastic-net Regression**	**RGLM**	**Ensemble KD**	**Ensemble Robust KD**
Correlation	0.349	0.35	0.18	0.765	0.063	0.701	0.721
Median Error	26.396	23.356	7.644	6.257	7.605	7.919	7.217

Figure 7.5: The association between the predicted biological age with DNAmAge suggests that the ensemble models are good estimates of a true biological trait.



7.6 Discussion

The estimated biological traits from the ensemble KD models have better accuracies and lower error compared to the KD models which provides support that the ensemble KD model can be used as an estimate for disease progression in DS studies. Though the biological trait estimated from the elastic-net regression has the highest accuracy and lowest error, it is only marginally different between DS and controls. We hypothesized the marginal effect may be attributed to the elastic-net algorithm where the primary objective is to optimize its predictability in chronological age. Consequently, it is too accurate in predicting chronological age resulting in a measurement that is more correlated with chronological age than disease progression by missing the subtle biological disruption caused by the condition.

We understand that a limitation of our study is in the pre-screen phase where we pre-selected a module of CpGs using WGCNA which may over emphasize the age acceleration effect in the DS subjects. However, we purposely chose a module of CpG’s that are associated with DS rather than chronological age because we wanted to capture CpG’s that may govern disease progression since CpGs that are related to disease may not have an effect on aging.

In all models evaluated, the estimated age accelerations are significantly higher in the DS subjects compared to the controls except for the Robust KD and the elastic-net regression which are only marginally significant. Nonetheless, the results consistently indicate that the DS subjects do experience an age accelerated effect that are evident in all the models assessed which are consistent with the results from Horvath and colleagues (Horvath et al., 2015).

7.7 Chapter 7 References

Devenny, D. A., Wegiel, J., Schupf, N., Jenkins, E., Zigman, W., Krinsky-McHale, S. J., & Silverman, W. P. (2005). Dementia of the Alzheimer's type and accelerated aging in Down syndrome. *Science's SAGE KE*, *2005*(14), dn1.

Horvath, S., Zhang, Y., Langfelder, P., Kahn, R. S., Boks, M. P., van Eijk, K., ... & Ophoff, R. A. (2012). Aging effects on DNA methylation modules in human brain and blood tissue. *Genome Biol*, *13*(10), R97.

Horvath, S., Garagnani, P., Bacalini, M. G., Pirazzini, C., Salvioli, S., Gentilini, D., ... & Franceschi, C. (2015). Accelerated epigenetic aging in Down syndrome. *Aging cell*, *14*(3), 491-495.

Jenkins, E. C., Ye, L., Gu, H., Ni, S. A., Duncan, C. J., Velinov, M., ... & Silverman, W. P. (2008). Increased "absence" of telomeres may indicate Alzheimer's disease/dementia status in older individuals with Down syndrome. *Neuroscience letters*, *440*(3), 340-343.

Kerkel, K., Schupf, N., Hatta, K., Pang, D., Salas, M., Kratz, A., ... & Jenkins, E. C. (2010). Altered DNA methylation in leukocytes with trisomy 21.

Kimura, M., Cao, X., Skurnick, J., Cody, M., Soteropoulos, P., & Aviv, A. (2005). Proliferation dynamics in cultured skin fibroblasts from Down syndrome subjects. *Free Radical Biology and Medicine*, *39*(3), 374-380.

Mather, K. A., Jorm, A. F., Parslow, R. A., & Christensen, H. (2011). Is telomere length a biomarker of aging? A review. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences*, *66*(2), 202-213.

Moran J. (2013). Aging and Down Syndrome. ndss: A Health & Well-Being Guidebook.

Parker, S. E., Mai, C. T., Canfield, M. A., Rickard, R., Wang, Y., Meyer, R. E., ... & Correa, A. (2010). Updated national birth prevalence estimates for selected birth defects in the United States, 2004–2006. *Birth Defects Research Part A: Clinical and Molecular Teratology*, *88*(12), 1008-1016.

Patterson, D., & Cabelof, D. C. (2012). Down syndrome as a model of DNA polymerase beta haploinsufficiency and accelerated aging. *Mechanisms of ageing and development*, *133*(4), 133-137.

Song, L., Langfelder, P., & Horvath, S. (2012). Comparison of co-expression measures: mutual information, correlation, and model based indices. *BMC bioinformatics*, *13*(1), 328.

Vaziri, H., Schächter, F., Uchida, I., Wei, L., Zhu, X., Effros, R., ... & Harley, C. B. (1993). Loss of telomeric DNA during aging of normal and trisomy 21 human lymphocytes. *American journal of human genetics*, *52*(4), 661.