GROWUPGROWOLD
HUMAN DEVELOPMENT, BIRTH TO DEATH

Moffitt & Caspi
Genes
Environment
Health
Behavior

Concept Paper Form

Provisional Paper Title: Comparison of feature selection methodologies and machine learning algorithms to identify a novel DNA methylation based estimator of telomere length
Proposing Author: Trevor Doherty and Dr Therese Murphy
Author's Email: therese.murphy@tudublin.ie
P.I. Sponsor: Professor Avshalom Caspi
Today's Date: 8/5/2021

Please describe your proposal in 2-3 pages with sufficient detail for helpful review.

Objective of the study:

The primary aim of this study is to apply a variety of machine learning methods to train a predictor of telomere length (TL) using epigenomic profiling data. It will provide a framework for identifying biological predictors of aging, uncovering biological insights into telomere biology and may lead to the identification of potential epigenomic biomarkers and/or therapeutic targets of aging and stress-related phenotypes like depression.

The primary Objectives of this study are:

1) Compare a range of feature selection methods and machine learning algorithms to train a predictor of TL based on DNA methylation in Dunedin sample cohort.
2) Develop a novel DNAm based estimator of TL and validate this estimator in two independent replication blood cohorts (n=192; n=178, respectively) collated in-house that have DNA methylation and TL measured.
3) Compare our DNAm based estimator of TL to a previously published estimator (1) and examine the overlap of CpG Sites between the two predictors.
4) Examine the relationship between a DNAm based estimator of TL and other DNAm based markers (e.g. DNAm Age (2) and Epismoker score (3) using Pearson's correlation coefficient.
5) Explore the predictability of a DNAm based signature for TL change overtime.

Data analysis methods:

Develop a novel DNAm based estimator of TL

Review of learning algorithms used in DNA-methylation-based estimators

DNA methylation-based machine learning studies and their utility to estimate and predict a range of quantitative traits including chronological age, telomere length (TL), epigenetic smoking scores and body mass index (BMI) (4) has increasingly become apparent. Feature selection methods are typically employed to reduce the extremely high dimensionality of input datasets and, in conjunction with a learning algorithm, estimate the quantitative trait of interest. We reviewed the literature to ascertain feature selection and learning algorithms commonly used in the epigenomics field. Due to the popularity in the published literature of elastic net penalised regression for epigenetic aging signatures, this approach will form one of the baselines in our study. Many studies also utilise some form of initial feature selection in advance of applying methods such as elastic net, therefore we will also investigate applying a range of common filters methods as an initial step. These will include, association tests corrected for multiple testing (false discovery rate (FDR) < 0.05), thresholding of correlation between CpGs and TL using Pearson's r, and variance-based filtering. Several comparative studies demonstrated that Support Vector Machines (SVR) performed well as an alternative to elastic net and other regression-based approaches, therefore we will also explore this method in conjunction with a range of filter methods. Additionally, tree-based methods that utilise inherent feature ranking such as random forest and gradient boosting models will be investigated. Combinations of feature selection methods (e.g. the intersection of top N features) and stacked models may also be explored.

Experiment 1 – Baseline

Given its common usage, we will utilise elastic net regression as a baseline model for TL prediction. The Dunedin dataset will be used for model training. Five-fold cross-validation will be applied to the data set in order to obtain optimal model hyperparameter settings. With the discovered settings, an elastic net model will be constructed on the Dunedin data and used to generate predictions on the 2 independent test data sets. The mean absolute error (MAE), Root Mean Square Error (RMSE), mean absolute percentage error (MAPE) and Pearson's r will be reported as measures of model generalisation performance.

Experiment 2 - Filter Methods with Elastic Net

Filter methods to be investigated include the use of Pearson's correlation coefficient, F-test with FDR threshold, variance-based filtering and the Relief algorithm. The Dunedin data set will be used for the purposes of training. To compare filter feature selection methods, MAE will be used as a measure of generalisation performance. Briefly, the Dunedin data set will be split into train and test data sets in 80%:20% proportions i.e. $DS1_{train}$ and $DS1_{test}$ respectively. Each filter method will be applied wholly to $DS1_{train}$, yielding a feature subset. Using only these features on $DS1_{train}$, 5-fold cross-validation will be conducted for hyperparameter tuning of an elastic net model.

Next, an elastic net model is constructed using $DS1_{train}$ (with the discovered feature subset and best parameters). This model is then used to generate predictions for $DS1_{test}$ and calculate an unbiased generalisation error. To protect against a fortuitous split of the train and test sets, 5 different train/test splits of the Dunedin data set are used with the described process applied to each (see Fig. 1). The final output is a mean MAE value ± standard deviation for the 5 disjoint test splits for each investigated filter method. The model with the lowest MAE corresponds to the best feature selection method. This evaluation strategy will identify the most promising feature filter method. In addition to the unbiased measure of generalisation performance already calculated using the Dunedin data set, we will extend our model evaluation to the 2 independent test sets. To obtain the feature subset for use with the independent test sets, the identified best feature selection method will be applied to the full Dunedin data set, returning an N feature subset. With these N features, 5-fold cross-validation is conducted on the full Dunedin data set, to find a good set of model hyperparameters. With the N features and best parameters identified, a model can be built using the entire Dunedin data set and used to generate predictions and calculate performance for the 2 independent test sets.

Experiment 3 - Filter Methods with Support Vector Regression
In addition to investigating filter methods with the elastic net learning algorithm (Experiment 2), we also explore the use of these filters with the SVR algorithm. Given the computational requirements of SVR, we apply this algorithm over a range of top ranked features from each filter method e.g. [100, 500, 1000, 2000, 3000, 4000, 5000]. Graphs of the number of features vs. performance metrics will be assessed to identify any obvious trends and optimal performance points. The same training, cross-validation and testing methodology will be used as described in Experiment 2.

Experiment 4 - Filter Methods with Tree-based Learning Algorithms
Having investigated filter methods with elastic net and SVR, we also explore the use of these filters with commonly used machine learning algorithms (e.g random forest regression (RFR) and gradient boosting (XgBoost)). These methods are computationally expensive, therefore we will implement these algorithms over a range of top ranked features from each filter method in a similar manner to that undertaken in Experiment 3. The same training, cross-validation and testing methodology will be used as described in Experiment 2

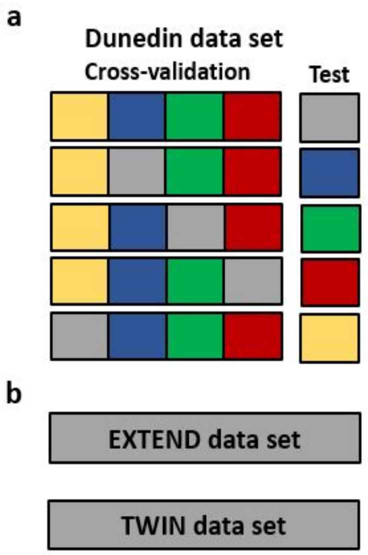

Figure 1: **a**. Over 5 runs, a feature selection method is applied to 4/5 of the Dunedin data set, followed by cross-validation on same to tune hyperparameters. The reduced feature set and best parameters on each run are used to make predictions on the final 1/5 of the data set (Test). The best feature selection method is chosen based on the average of the 5 test sets. **b**. Two independent data sets are used to further assess the generalisation performance. In this case, the feature selection method is applied to the whole Dunedin data set to acquire the reduced feature subset. With this feature set, cross-validation is conducted on the whole Dunedin data set to obtain the best set of hyperparameters. Finally, a model is built using the entire Dunedin data with the reduced feature set and best hyperparameters. This model is used to predict on both independent data sets, yielding a measure of unbiased generalisation performance on each data set.

Examining the relationship between DNAm based TL estimates and other DNAm based markers.
To assess the effect of other DNAm based markers ((e.g. DNAm Age and Epismoker score).
on DNAm TL estimator in blood, we will calculate DNAm Age (2) and Epismoker score (3) for all 3 datasets and examine their correlation using Pearson's r.

Explore the predictability of a DNAm based signature for TL change overtime
First, an estimate of TL change (ΔTL) between age 26 and 38 will be calculated by regressing TL age 38 on TL at age 26 (while controlling for confounders). Next, applying the same methodology outlined above we will examine the predictability of a DNAm bases estimator of ΔTL (shortening or lengthening of TL) in the Dunedin longitudinal cohort.

<u>Variables needed at which ages:</u>

All variables are needed at both age 26 and 38 unless otherwise stated.

- Telomere length (T/S ratio per diploid genome)
- Normalised DNA methylation values

- Gender
- Measured cellular composition measures
- Sentrix ID/Chip ID
- Smoking

Significance of the Study (for theory, research methods or clinical practice):

Telomeres are DNA repeat structures at the ends of chromosomes, which have a crucial role in maintaining genomic stability. Furthermore, they play a critical role in regulating cellular replication, whereby TL gradually shortens due to the failure of DNA polymerase to fully replicate the 3' end of the DNA strand, thus accruing cellular damage (5). Inter-individual variation in mean TL has been associated with cancer and several age-associated diseases(5) and TL has emerged as a promising biomarker for biological age(6). In recent years an association between TL and DNA methylation has been hypothesised. For example, DNA methyltransferases (the enzyme that maintain DNA methylation) are known to control TL in mammalian cells(7), highlighting a link between DNA methylation and TL homeostasis. Recently, a DNA-based TL estimator was created based on 140 CpGs using a regression based machine learning approach (i.e elastic net) (1). Our study wishes to build on this previous study by evaluating a range of feature selection methodologies and machine learning algorithms to identify a novel DNA methylation based estimator of TL. By the end of the project we will have a DNAm based estimator of TL than can be further evaluated to a) examine associations with features of human aging-related traits and behaviors, and b) to gain a better understanding of the molecular underpinnings of such associations. Moreover, we will have a robust methodology utilising machine learning algorithms, which could be applied to other biological markers and disease phenotypes, to examine their relationship with DNA methylation.

References cited:

1. Lu AT, Seeboth A, Tsai PC, Sun D, Quach A, Reiner AP, et al. DNA methylation-based estimator of telomere length. Aging (Albany NY). 2019;11(16):5895-923.

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

3. Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. Epigenomics. 2019;11(13):1469-86.

4. Hamilton OKL, Zhang Q, McRae AF, Walker RM, Morris SW, Redmond P, et al. An epigenetic score for BMI based on DNA methylation correlates with poor physical health and major disease in the Lothian Birth Cohort. Int J Obes (Lond). 2019;43(9):1795-802.

5. Codd V, Nelson CP, Albrecht E, Mangino M, Deelen J, Buxton JL, et al. Identification of seven loci affecting mean telomere length and their association with disease. Nature genetics. 2013;45(4):422-7, 7e1-2.

6. Benetos A, Okuda K, Lajemi M, Kimura M, Thomas F, Skurnick J, et al. Telomere length as an indicator of biological aging: the gender effect and relation with pulse pressure and pulse wave velocity. Hypertension. 2001;37(2 Pt 2):381-5.

7. Gonzalo S, Jaco I, Fraga MF, Chen T, Li E, Esteller M, et al. DNA methyltransferases control telomere length and telomere recombination in mammalian cells. Nature cell biology. 2006;8(4):416-24.

Data Security Agreement

Provisional Paper Title: Comparison of feature selection methodologies and machine learning algorithms to identify a novel DNA methylation based estimator of telomere length

Proposing Author: Trevor Doherty and Dr Therese Murphy

Today's Date: 8/5/2021

☒	I am current on Human Subjects Training (CITI (www.citiprogram.org) or equivalent)
☒	My project is covered by the Duke ethics committee OR I have /will obtain ethical approval from my home institution.
☒	I will treat all data as "restricted" and store in a secure fashion. My computer or laptop is: a) encrypted (recommended programs are FileVault2 for Macs, and Bitlocker for Windows machines) b) password-protected c) configured to lock-out after 15 minutes of inactivity AND d) has an antivirus client installed as well as being patched regularly.
☒	I will not "sync" the data to a mobile device.
☒	In the event that my laptop with data on it is lost, stolen or hacked, I will immediately contact Moffitt or Caspi.
☒	I will not share the data with anyone, including my students or other collaborators not specifically listed on this concept paper.
☒	I will not post data online or submit the data file to a journal for them to post. *Some journals are now requesting the data file as part of the manuscript submission process. Study participants have not given informed consent for unrestricted open access, so we have a managed-access process. Speak to Temi or Avshalom for strategies for achieving compliance with data-sharing policies of journals.*
☒	I will delete all data files from my computer after the project is complete. Collaborators and trainees may not take a data file away from the office. This data remains the property of the Study and cannot be used for further analyses without an approved concept paper for new analyses.
☒	I have read the Data Use Guidelines and agree to follow the instructions.

Signature: Therese Murphy