Introduction

- Chronological age is a risk-factor for non-communicable diseases including cardiovascular disease
- Biological ageing which is the aggregate effect of cellular and biochemical processes… could better reflect the individual risk of pathology
- There have been many biomarkers used…
- Many biological clocks have been constructed from these biomarkers… Hannum, Horvath, Levine methylation, since then other multi-omics clocks have come out metabolomics, iAge, proteomics
- None have combined all three platforms, it remains unclear what methods are best due to their interconnectedness.

Aims

- Determine whether a combination of omics platforms incurs better prediction of chronological age i.e., biological clock and the best method
- Is it a better surrogate measure for CACS then FRS

Methods

Cohort information

The dataset used is derived from BioHEART-CT biobank – an Australian prospective, longitudinal cohort study of patients who had CT coronary angiograms with coronary artery calcium score (CACS) (Kott et al., 2019). The cohort contained 969 adult participants with 540 males (56%). The mean age was 60.8 (SD=12.2, range: 21-94 years). Most of participants (86.8%) were of European ethnicity.

Multi-omics datasets were available as per FIGURE. The dataset contained metabolomics, proteomics, lipidomics and CyTOF data which was normalised by hierarchical remove-unwanted-variation (hRUV) method (Kim et al., 2021).

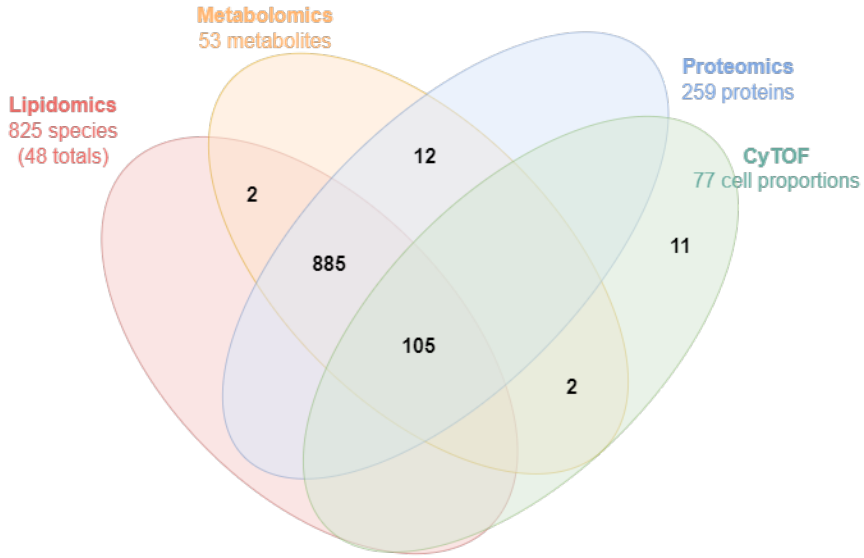

Figure 1 - Available patients for various combination of omics platforms.

Biological age modelling

All -omics datasets were split into independent training (70%) and test (30%) sets. All modelling was performed on the independent sets and out-of-sample performance was validated on the held-out test sets. All analysis took place R ver. 4.1.0. Five common machine learning methods were chosen as candidates due to their previous use; elastic net, PCA regression, random forest, deep neural network (DNN), autoencoder regression (Acharjee, 2012).

In all modelling situations, the chronological age was regressed on either a subset of analytes or a dimensionally reduced representation of the assay data. The predictions made by the resulting model were considered the biological age. All of the modelling was performed on metabolomics, proteomics, lipidomics, CyTOF separately. The modelling was then repeated on a concatenated set of lipidomics, proteomics and metabolomics. CyTOF was omitted from this combined analysis because the intersection was insufficient as in **FIGURE**

Elastic net regression was modelled using "glmnet" (Friedman et al., 2021). 5-fold repeated cross validation (n=5) using "caret" was performed to find optimal hyperparameters (Kuhn et al., 2022). This random search for *alpha* (describes the proportion of ridge and LASSO penalties) and *lambda* (severity of the penalty for large coefficients with small predictive value) hyperparameters optimised for minimum cross-validated RMSE.

PCA regression involved a multi-stage process. Firstly, single value decomposition of the assay matrix was performed to find principal components (PC). The 'elbow' of the scree plot indicated an abrupt drop in explained variance by the subsequent PCs and these PCs were dropped. Chronological age was then regressed on the PCs before the 'elbow' using multiple linear regression (MLR).

Random forest was modelled using Breiman's algorithm in "randomForest" (Cutler & Wiener, 2022). 5-fold repeated cross validation (n=5) using "caret" was used to tune hyperparameter *mtry* (number of randomly selected variables to split on at each branch) (Kuhn et al., 2022).

Neural network architecture consisted of two hidden fully connected layers (256, 64) using the ReLU activation function. The model training used an objective function of mean absolute error, an Adam optimiser (learning rate=0.001) and "keras" (Kalinowski et al., 2022).

Autoencoder regression was performed using 'h2o' (LeDell et al., 2022). The architecture consisted of a single hidden bottleneck layer (250) and tanh activation function due to the comparatively low dimensionality of the data compared to typical autoencoder applications. The encodings stored within the deep features of the bottleneck layer were used to train a MLR with chronological age as the regressand.

Validation of biological age models

Bootstrapped (n=200) confidence intervals for the R^2 and mean absolute error (MAE) were calculated for the fit on chronological age on the out-of-sample test data. This was performed separately for each model and assay.

Calculation of biological age acceleration

The best model was used to calculate the biological age – defined as the fitted value of chronological age regressed on assay. The biological age acceleration (bAA) was defined as the residuals of biological age regressed on chronological age. This adjusts for chronological age in a similar way to the original DNA methylation clock procedure (Horvath, 2013).

Risk factors of biological age acceleration

The bAA was then regressed on hypertension, diabetes mellitus, angina, rheumatoid arthritis, gout,

Benjamini & Hochberg

Datasets were split into a training and testing set (70:30) and the test set was held out during training to prevent over-fitting.

To benchmark different machine learning methods, according to Aim #1, five common machine-learning methods were chosen; Elastic Net regression, Principal Component regression, Random Forest, XGBoost and Deep Neural Network.

To compare the efficacy of models, bootstrapped 95% confidence intervals (n=200) of R2 score and MAE were calculated for the fit of predicted age versus chronological age on the test data.

The best model was then used on the test data to calculate an age acceleration by taking the residuals of predicted age versus chronological age. Furthermore, a measure for resiliency was calculated by taking the residuals of CACS percentile (which is a score adjusted for age and sex) regressed on FRS. Resiliency was then plotted again age acceleration and age to determine whether there was a relationship as per Aim #2.

Discussion

Combined Platform was Better

Elastic Regression vs Other Methods

- Elastic regression performed best across all data types, indicating that it translates well to high p to n omics data. It's a convex combination of both L1 and L2 regularization that penalises excess covariates with minimal contribution to the outcome variable and encourages the final model towards parsimony.
- It's possible PCR performed poorly because the direction of greatest explained variance doesn't necessarily contain the most information about the outcome of interest. Since we limit the number of latent variables used, it's possible that important information in unused axes were crucial to the age prediction.
- Importantly deep neural networks did not perform well possibly due to a number of factors: 1) they need far more data than ML counterparts to avoid overfitting 2) high dimensionality means far more irrelevant features and mass spectrometry known to be very noisy
- Tried autoencoder which would perform dimensional reduction, feed the encoded layer of deep features into a MLR. Certain features may have a lot of influence on

other features but importantly do not correspond directly with the age and therefore don't explain it well.

Reduced Performance Compared to Other Clocks

- See Google Docs

Predictive Ability of CACS

-

Chronological age is a well-known risk factor for many chronic conditions however, measuring age based on the time elapsed is somewhat arbitrary.

A more clinically relevant measurement, would capture the aggregate effect of cellular & biochemical processes in our body produced by genetic and environmental factors that translates to physiological impairment. We term this biological age.

The biomarkers that have been used over time to predict biological age have evolved from physical e.g. declining cognition, to clinical e.g. high LDL, to cellular e.g. telomere attrition and more recently, molecular with multi-omics data.

In more concrete terms, the biological age is predicted by regressing chronological age on a series of biomarkers. The biological age acceleration - used in later analysis - is defined as the residuals. So a negative residual would indicate slower aging and whereas a positive residual would indicate faster aging.

Clinical datasets introduce a unique challenge due to the cohort heterogeneity that arises from the divergent characteristics of subjects. This is exacerbated by the use of high throughput multi-omics data which is characteristically noisy. Whilst there are studies that use traditional ML methods of elastic regression on -omics data, there hasn't been a comparison of different methods on the same dataset. Therefore, my first aim is to:

1. Determine the best combination of linear/non-linear machine-learning method and omics platform/s for estimating chronological age.

Now, biological age has a number of applications but I only have time to dive into one.

Biological age, in theory, allows you to evaluate individual risk rather than population risk, particularly for risk scores that rely heavily on chronological age; for example the Framingham Risk Score for cardiovascular disease. There is a sizeable resilient group with high FRS and low to no coronary artery calcification consistently reported clinically. This reveals the shortfall of FRS being a measure of cohort rather than individual risk. Therefore my second aim is to:

1. Determine whether there is a relationship between biological age acceleration and resiliency that could help explain patients that FRS (which heavily depends on chronological age) can not.

```
Analysis of Variance Table

Response: pred
                     Df Sum Sq Mean Sq   F value    Pr(>F)
age                   1  55724   55724 2457.9192 < 2.2e-16 ***
gender                1      8       8    0.3455  0.556846
ethcat                1     87      87    3.8443  0.050223 .
cvhx_htn              1     58      58    2.5434  0.111108
cvhx_dm               1    373     373   16.4403 5.454e-05 ***
bmi                   1    116     116    5.0948  0.024235 *
signif_smok           1      0       0    0.0000  0.996996
drinking_status       1     35      35    1.5593  0.212088
fh_ihd                1     13      13    0.5615  0.453832
fh_clottingdisorders  1      0       0    0.0006  0.980910
cvhx_hf               1     86      86    3.8034  0.051457 .
cvhx_angina           1    136     136    6.0035  0.014467 *
cvhx_mi               1      1       1    0.0372  0.847194
cvhx_rhythm_af        1     69      69    3.0469  0.081231 .
mhx_arthritis_ra      1    139     139    6.1306  0.013468 *
mhx_arthritis_osteo   1      0       0    0.0140  0.905722
mhx_arthritis_gout    1     90      90    3.9790  0.046370 *
mhx_osteoporosis      1     68      68    3.0051  0.083344 .
mhx_stroke            1    115     115    5.0619  0.024697 *
mhx_pad               1      0       0    0.0133  0.908383
mhx_dvt_pe            1    166     166    7.3307  0.006907 **
mhx_kidney            1    172     172    7.5665  0.006065 **
Residuals           903  20472      23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```