|  |  |
| --- | --- |
| Good afternoon judges and distinguished guests. I’ll speak to you today, about machine-learning modelling of biological age and an interesting application on atherosclerotic disease.  **NEXT SLIDE**  Chronological age is a well known risk factor for many chronic conditions however, measuring age based on the time elapsed is somewhat arbitrary.  **NEXT SLIDE** | 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.  **NEXT SLIDE**  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.  **NEXT SLIDE** | 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.  **NEXT SLIDE**  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: |
| Determine the best combination of linear/non-linear machine-learning method and omics platform/s for estimating chronological age.  **NEXT SLIDE**  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: |
| 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.  **NEXT SLIDE**  Now for Methods.  **NEXT SLIDE** | The multi-omics dataset used was acquired from the BioHEART-CT, a prospective cohort study of patients with suspected coronary artery disease. It contains Metabolomics, Proteomics, Lipidomics and CyTOF data which was already normalised via a pairwise hierarchical method.  Datasets were split into a training and testing set (70:30) and the test set was held out during training to prevent over-fitting. |
| I benchmarked different machine learning methods, according to Aim #1. Four common machine-learning methods were chosen; Elastic Net regression, Principal Component regression, Random Forest, XGBoost and Deep Neural Network.  **NEXT SLIDE**  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. | **NEXT SLIDE**  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 score was then plotted again age acceleration and age to determine whether there was a relationship as per Aim #2. |
| **NEXT SLIDE**  This is the full pipeline.  **NEXT SLIDE**  Now to talk a little bit about my results and discussion.  **NEXT SLIDE**  This was my full results which you can take a closer look at in the appendix. | **NEXT SLIDE**  The comparative performance plot here shows R2 scores across assays and ML methods.  For Aim #1, Elastic regression performed best across almost all data types, indicating that it translates well to high p to n omics data. This is likely because the convex combination of both L1 and L2 regularization penalises excess covariates with minimal contribution to the outcome variable and encourages the final model towards parsimony. |
| In comparison, PCR similarly performs dimensional reduction, by regressing on a subset of latent variables that maximally explains variance - but importantly, doesn’t necessarily contain the most amount of information about the outcome. This could be why it performed the worst.  Importantly, deep neural networks did not perform well despite their popularity in the dogma. It’s known that DNNs need far more data compared to traditional ML modelling to avoid overfitting on noise so this could explain it’s poor performance in our small dataset. | Our best model was Elastic Net regression on a combination of Proteomics, Metabolomics and Lipidomics (which we termed PML) with an R2 score of 0.59.  **NEXT SLIDE**  Here we are now comparing our best model with biological clocks in literature. Unfortunately, our best model which was Elastic Net regression on a combination of Proteomics, Metabolomics and Lipidomics (which we termed PML) still performed worse than other -omics models in literature. |
| This could be because the present dataset has far less samples diminishing the model’s out-of-sample predictive power.  It’s also worth noting that the panel size i.e. feature set measured, for the other datasets in literature were much larger and closer to un-targeted. It’s possible that features which would have been major contributors to the biological age prediction were not present in the current dataset. | One strategy to overcome this in the future could be to source an untargeted dataset with a larger feature set. A problem with that strategy is that there may be far more predictors than samples in which case, I could use an auto-encoder or tSNE pre-processing to dimensionally reduce it prior to regression.  **NEXT SLIDE**  As for Aim #2, we first plotted the PML age against Chronological age to find the age acceleration which is the residual. As expected, the coronary artery calcification tends to be greater with age.  **NEXT SLIDE** |
| Next we regressed the CACS percentile against the cardiovascular FRS. Again the residual here represents the resilience score.  **NEXT SLIDE**  Finally, we took the resiliency score and regressed it on age acceleration and found the trend to oppose our initial expectation. This seems to suggest that resiliency increases with age acceleration.  **NEXT SLIDE** | It’s possible that this is because both resiliency and age acceleration are correlated with chronological age which could be the cause of the paradoxical flipped relationship. It’s also possible that the unintended relationship is coerced by the large group of chronologically young individuals who naturally have low CACS and a high resiliency score, under this definition. |
| An alternative strategy would be to re-define the resiliency problem by using the age acceleration instead of chronological age to calculated an adjusted FRS. The CACS percentile could then be regressed on either the FRS or the adjusted-FRS to compare the goodness of fit. If the adjusted-FRS produced a better goodness of fit, it would be “explaining” the resilient individuals better - whom are inherently outliers of the original FRS model.  **NEXT SLIDE**  **NEXT SLIDE** | Therefore, elastic net regression still outperforms other models. A larger feature set and sample size with appropriate dimensional reduction would produce a benchmark more representative of contemporary models.  A paradoxical relationship was found between age acceleration and resilience. Evidently, a re-framing of the problem with adjusted-FRS is necessary to circumvent statistical anomalies arising due to multi-covariance. |