Title: Machine Learning for the Development of Methylation Risk Score Predictors

Authors: Andres Cardenas, Dennis Khodasevich, Nina Holland, Lars van der Laan

Background: DNA methylation (DNAm) provides a window to characterize the impacts of environmental exposures and the biological aging process. Epigenetic clocks are often trained on DNAm using penalized regression of CpG sites, but recent evidence suggests potential benefits of training epigenetic predictors on principal components.

Methodology/Findings: We developed a pipeline to simultaneously train three epigenetic predictors; a traditional CpG Clock, a PCA Clock, and a SuperLearner PCA Clock (SL PCA). We gathered publicly available DNAm datasets to generate i) a novel childhood epigenetic clock, ii) a reconstructed Hannum adult blood clock, and iii) as a proof of concept, a predictor of polybrominated biphenyl exposure using the three developmental methodologies. We used correlation coefficients and median absolute error to assess fit between predicted and observed measures, as well as agreement between duplicates. The SL PCA clocks improved fit with observed phenotypes relative to the PCA clocks or CpG clocks across several datasets. We found evidence for higher agreement between duplicate samples run on alternate DNAm arrays when using SL PCA clocks relative to traditional methods. Analyses examining associations between relevant exposures and epigenetic age acceleration (EAA) produced more precise effect estimates when using predictions derived from SL PCA clocks.

Conclusions: We introduce a novel method for the development of DNAm-based predictors that combines the improved reliability conferred by training on principal components with advanced ensemble-based machine learning. Coupling SuperLearner with PCA in the predictor development process may be especially relevant for studies with longitudinal designs utilizing multiple array types, as well as for the development of predictors of more complex phenotypic traits.