Predicting onset, progression, and clinical subtypes of Alzheimer's and Parkinson's disease from genomic and longitudinal clinical data

Faraz Faghri^{1,2,3}, Vipul Satone², Rachneet Kaur², Sayed Hadi Hashemi², Hampton Leonard¹, Sonja W. Scholz¹, Mike A. Nalls¹, Andrew B. Singleton¹, Roy H. Campbell²

¹ National Institute on Aging, National Institutes of Health
² Department of Computer Science, University of Illinois at Urbana-Champaign
³ Mayo Clinic

Background and Goal

Parkinson's disease, Alzheimer's Disease and Related Dementia are categories of **aging disorders** with **complex and heterogeneous symptomatology** where accurate subtype classification, diagnosis, and prognosis, has the potential to be of great benefit for researchers and clinical practice. We **describe and classify subtypes** of phenotypic changes in patients with these aging disorders over months. We also create **accurate prediction for the diagnosis and prognosis**.

Data

We used machine learning methods on comprehensive, longitudinal clinical data from the Parkinson Disease Progression Marker Initiative (PPMI) (n=328 cases) and Alzheimer's Disease Neuroimaging Initiative (ADNI) (n=248). The resulting models were validated in an independent, clinically well-characterized cohort from the Parkinson Disease Biomarker Program (PDBP) (n=112 cases).

Analytical Pipeline

We used unsupervised and supervised machine learning approaches for subtype identification and prediction.

Unsupervised Learning: disorder types, subtypes, and interplay

Our analysis distinguished three distinct disease subtypes with highly predictable progression rates, corresponding to slow, moderate and fast disease progressors. We also developed an **embedding space** representing **patient's trajectory 7-years from diagnosis** in terms of **Sleep, Cognition, and Movement** disturbances.

Supervised learning: diagnosis, prognosis trajectory prediction

We achieved highly accurate projections of disease progression four years after initial diagnosis with average Area Under the Curve of (95% CI: 0.96 ± 0.01).

Challenges and Lessons Learned

Data Modality (genetic, clinical, imaging, etc)	Notes
	<ul style="list-style-type: none">- Data curation takes significant time (~3 months), while the machine learning dev is quick (~1 week)- Data might be found useless due to significant missingness or data entry errors- Traditional methods (Bayesian, random forest, SVM, etc) work great on non-longitudinal- Longitudinal data requires new approaches, <i>data vectorization</i> and <i>LSTM</i> are among them
	<ul style="list-style-type: none">- Data curation doesn't take much time (~ 2 weeks), while the machine learning is both human and computation intensive (~2 months)- <i>Boosting methods</i> work great while <i>deep learning method</i> requires much larger samples sizes- <i>Parameter tuning</i> is computation intensive and takes significant time
	<ul style="list-style-type: none">- Both data curation and machine learning dev are time consuming, a project usually takes 6-12 months- No straightforward machine learning method yet, still an open problem- For <i>end-to-end deep learning techniques</i>, 3D brain images is one of the most challenging data modes