**Spring 2022 Seminar Series**

# Explainable machine learning aggregates polygenic risk scores and electronic health records for Alzheimer's disease prediction

**Abstract:** Background: Alzheimer's disease (AD) is the most common late-onset neurodegenerative disease. Identifying individuals at increased risk of developing AD is important for early intervention. Risk prediction models are typically based on a limited number of predictors, possibly with sub-optimal performance. Here, we explore an explainable machine learning (ML) framework, XGBoost and SHapley Additive exPlanations (SHAP) values, for AD risk prediction, which can handle a large number of predictors and output the impact and importance of each predictor. Method: We developed an XGBoost model that aggregates polygenic risk scores (PRSs), baseline individual characteristics (e.g., non-genetic factors), and information from electronic health records for predicting incident AD. The PRSs were derived using summary statistics from genome-wide association studies in the Alzheimer's Disease Genetics Consortium (ADGC) dataset (n = 19,918). The model was applied to 457,936 white participants in UK Biobank to predict development of AD within 10 years after the baseline visit (n = 2,177 developed AD). We further used SHAP values to explain the relative information in model predictors. Result: For participants of age 40 and older, the area under the receiver operating characteristic curve (AUC) for AD risk prediction was over 0.880. PRSs ranked second to age (the best predictor) in feature importance. For subjects of age 65 and above, PRSs for AD were the most important features. Our ML model not only identified traditional risk factors for AD, such as age, education, income, body mass index, diabetes, and blood pressure, but also identified predictors from electronic health records that are not typically considered in traditional prediction models, including urinary tract infection, syncope and collapse, chest pain, disorientation and hypercholesterolaemia, for developing AD. Furthermore, SHAP values aided the ranking of feature importance and model explanation. Conclusion: Our ML model improves the accuracy of AD risk prediction by efficiently exploring numerous predictors. PRSs play the most important role in developing AD in individuals of age 65 and older. In application, the model also identified novel feature patterns for AD.

**About the Speaker:** Dr. X. Raymond Gao is an Associate Professor at the Ohio State University. He has a joint appointment in Ophthalmology, Biomedical Informatics, and Human Genetics. Dr. Gao is an experienced statistical geneticist and a leader of ocular genomics research for multiple National Eye Institute funded projects on glaucoma genetics, steroid-induced ocular hypertension and ocular biostatistics and genetic analysis. He is also a principal investigator of a National Institute on Aging funded project, genetic basis of age-at-onset of Alzheimer disease. His research program involves interdisciplinary studies in genetics, statistics and machine learning. His lab also received funding for developing AI for glaucoma. His research goals are to advance our understanding of the genetic architecture of complex human diseases and traits and to improve our ability to predict disease, target prevention to high-risk individuals and tailor treatment based on individual genomic differences.

**X. Raymond Gao, PhD**
**Associate Professor**
**OSU, Ophthalmology, Biomedical Informatics, and Human Genetics**
**Friday, January 14th, 11:00am-12:00pm**
**Carmen Zoom**