Machine Learning in Neuroimaging: from Application to Interpretation

Elvisha Dhamala Organizer
Yale University
New Haven, Connecticut
USA

Avram Holmes Co-Organizer
Departments of Psychology and Psychiatry, Yale University
New Haven, Connecticut
USA

Overview

The application of machine learning in neuroimaging over the past two decades has enabled the identification of individualised biomarkers of health and disease. Predictive models have successfully been implemented to capture imaging signatures of typical development and aging, cognitive and behavioral traits, and psychiatric and neurodegenerative illnesses. While these findings have undoubtedly advanced our understanding of brain-behavior relationships, the clinical utility of these models is highly debated because of their limited ability to identify the specific neurobiological processes underlying those relationships. This limitation is often due to the trade-off between model accuracy (the model's ability to make successful predictions) and interpretability (our ability to understand how the model works). Accuracy typically comes at the cost of interpretability, but interpretability is irrelevant if models are inaccurate.

In this symposium, we will address challenges of applying machine learning in neuroimaging as they pertain to model accuracy and interpretability. We will begin by discussing the lack of feature weight reliability in prediction models and providing strategies that can be implemented to increase reliability of the feature weights and subsequently the interpretability of the models. Next, we will explore how implementing these strategies can influence the relationship between feature weight reliability and prediction accuracy when modelling behaviors across three separate domains: cognition, personality, and mental health. Finally, we will outline how these strategies can be leveraged to develop models to capture shared and unique brain-behavior relationships across races, ethnicities, age groups, and sexes. This is a timely topic given the rapidly increasing interest within the human brain mapping to apply machine learning techniques to reveal fundamental insights into population-specific relationships between neurobiological features and behavioral phenotypes in health and disease.

Lecture 1: *Machine Learning Facilitates Generalizable Associations with Cognitive and Clinical Measures in Large-scale Developmental Neuroimaging Datasets*

Arielle Keller Presenter

Neuroimaging studies require increasingly large datasets to uncover reliable and generalizable brain-behavior associations. These large datasets require researchers to carefully balance the tradeoff between resource-heavy participant-level analyses that account for inter-individual heterogeneity in structure and function and group-level analyses that facilitate interpretability and generalizability. In this talk, I outline these tradeoffs and detail two examples of recent work using large-scale developmental neuroimaging datasets. In these examples, we leverage machine learning methodologies to achieve both participant-level detail and generalizable, interpretable findings. First, we demonstrate the utility of regularized non-negative matrix factorization to define reproducible and interpretable personalized functional brain networks in two large-scale datasets of youth (PNC and ABCD). Replicating findings across datasets, we show that individually-defined fronto-parietal network topography is associated with cognitive abilities in youth, and use machine learning to predict cognitive abilities in held-out participants' data. Second, we develop machine learning models to predict borderline personality disorder symptoms using whole-brain functional connectivity across two datasets (HCP-YA and HCP-D). We find that these models can successfully predict symptoms in unseen data from both adults and older adolescents. Furthermore, we find that functional connectivity patterns linked to borderline personality disorder symptoms are also those that undergo the most development in youth. Together, these findings demonstrate that machine learning approaches can be leveraged to generate interpretable and generalizable predictions across multiple large-scale developmental neuroimaging datasets.

Lecture 2: *Prediction accuracy and test-retest reliability of feature importance for behavioral prediction*
Leon Qi Rong Ooi Presenter

There is an increasing interest in neuroimaging-based behavioral prediction. Feature importance of the prediction models is usually extracted to understand the relationship between imaging features and behaviors (Finn, 2015). However, a recent study showed that feature importance of cognition prediction models had low test-retest reliability across data samples and suggested a trade-off between feature importance reliability and prediction accuracy (Tian, 2021). To quantitively explore the relationship between feature importance reliability and prediction accuracy, we used resting functional connectivity (rest-FC) to predict behavioral measures across three behavioral domains (cognition, personality, mental health) in the Adolescent Brain Cognitive Development (ABCD) study (Casey, 2018). We tested the feature importance reliability under multiple factors including interpretation methods, behavioral domains, and sample sizes. In this talk, we will first show the factors affecting feature importance reliability and suggest the best practice for obtaining reliable feature importance. We will then examine the empirical and mathematical relationship between feature importance reliability and prediction accuracy. Finally, we will discuss why reliable feature importance is not equivalent to correct feature importance and provide the general guidelines to obtain the correct feature importance.

Lecture 3: *Cross-ethnicity/race generalization failure of RSFC-based behavioral prediction and potential downstream consequences*
Jingwei Li Presenter

Algorithmic biases that favor majority populations pose a key challenge to the application of machine learning for precision medicine. In neuroimaging, there is growing interest in the prediction of behavioral phenotypes based on resting-state functional connectivity (RSFC). In that context, predictive models are typically built by capitalizing on large cohorts with mixed ethnic groups, in which the proportions of certain groups, e.g. African Americans (AA), are limited. Here, we investigated cross-ethnicity/race generalizability of the current, field-standard behavioral prediction approach using two large-scale public datasets from the United States. Specifically, we observed larger prediction errors in AA than white Americans (WA) for most behavioral measures using both the Human Connectome Project (HCP) and the Adolescent Brain Cognitive Development (ABCD) data. This prediction bias towards WA corresponded to more WA-like brain-behavior association patterns learned by the models. Looking into the direction of prediction errors, concerns can be raised if the machine-learning prediction

results would be uncritically used, in particular for the diagnosis of mental disorders. For example, social support measures were more overpredicted for AA than WA, whereas social distress measures such as Perceived Rejection were more underpredicted for AA than WA.
Furthermore, African pre-adolescent participants suffered from more overpredicted social problems, rule-breaking and aggressive behaviors compared to white participants. Effects of the training populations were also studied by comparing predictive models trained specifically on AA, specifically on WA, or on a mixture of AA and WA with equal sizes. Although specific training on AA slightly helped to reduce the biases against AA, most behavioral measures still exhibited larger prediction errors in AA than WA. Other possible sources of the biases such as neuroimaging preprocessing (e.g., brain templates and functional atlases) and the design of behavioral measures need to be examined in the future.

Lecture 4: *Leveraging machine learning in neuroimaging to capture population-specific brain-behavior relationships*

Elvisha Dhamala Presenter

Machine learning is increasingly being used in the field of human brain mapping to understand brain-behavior relationships. Prior research has typically focused on the application or interpretation of predictive models, but seldom both. Several studies have evaluated differences in prediction accuracy as a function of the machine learning algorithms applied, the neurobiological features used, or the behaviors predicted. Others have investigated the effects of algorithmic choice and feature weight transformations on model interpretation. However, relatively little is known about the extent to which these predictive models perform and generalize across different populations, or the unique population-specific features that drive those predictions.

In this talk, I will highlight how population-specific predictive models can be used to capture accurate and interpretable brain-behavior relationships. First, I will present results from two sets of experiments examining how age- and sex- specific predictive models can be employed to (a) estimate differences in the strength of brain-behavior relationships across ages and sexes, and (b) identify age- and sex- specific neural correlates of behavior. Next, I will argue that rather than regressing out the effects of age, sex, and other demographic variables, as is standard practice in the field, we should use population-specific models to quantify the effects of those variables on the relationships being analyzed. Finally, I will provide insights into how a similar approach can be applied to better understand the behavioral effects of neurological and psychiatric illnesses in distinct demographic and clinical populations.