Translating Big Data Imaging Genomics Findings to the Individual: Prediction of Risks and Outcomes in Neuropsychiatric Illnesses

Peter Kochunov
Department of Psychiatry and Behavioral Sciences, University of Texas Health Science Center at Houston and UT Health Houston School of Behavioral Health Sciences, Houston, TX, USA

Li Shen
Department of Biostatistics, Epidemiology and Informatics
Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
Email: li.shen@pennmedicine.upenn.edu

Zhongming Zhao
Center for Precision Health, McWilliams School of Biomedical Informatics
University of Texas Health Science, Center at Houston, Houston, TX, USA

Paul M. Thompson
USC Mark and Mary Stevens Neuroimaging and Informatics Institute
Keck School of Medicine, University of South California, Los Angeles, CA, USA
Email: pthomp@usc.edu

This PSB 2025 session is focused on opportunities, challenges and solutions for translating Big Data Imaging Genomic findings toward powering decision making in personalized medicine and guiding individual clinical decisions. It combines many of the scientific directions that are of interest to PSB members including Big Data analyses, pattern recognition, machine learning and AI, electronic health records and others.

1. Introduction

National and international scientific efforts are expanding toward collection, sharing and analyses of large and inclusive epidemiological and illness-focused datasets that combine genetic, imaging, metabolic and electronic health records (EHRs) data to enable examination of the contribution of genetic, environmental and interventional factors to human illness and health. High-resolution neuroimaging (~10^{4-6} voxels), genetic (10^{6-8} single nucleotide polymorphic variants (SNPs)) and EHRs (~10^{2-5} structured features + clinical notes) per individual are available in statistically powerful (N=10^{3-5}) epidemiological and disorder-focused samples. This also leads to major challenges on collection, sharing and homogenization of data, including how to identify reproducible signatures of complex polygenic illnesses. Research findings in such illnesses, e.g., neuropsychiatric, neurodegenerative, metabolic and other complex disorders, have historically suffered from a substantial variability and heterogeneity both within and across disorders - including genetics, environmental risk factors, mean age of onset, symptom presentations, treatment response, and long-term prognosis. Sources of heterogeneity have long remained a challenge to clinicians and scientists and have contributed to a surprisingly poor reproducibility

and difficulty in translating research findings to personalized risk assessments that can guide clinical decisions.

Presentations in this session demonstrate how Big Data collaborations such as IBM Watson Health, UK Biobank (UKBB), Enhancing Neuro Imaging Genetics through Meta Analyses (ENIGMA), the Human Connectome Project (HCP), Alzheimer's Diseases Neuroimaging Initiative (ADNI), Psychiatric Genetics Consortium (PGC), Penn Medicine EHRs and others have enabled novel principled approaches to reduce false positive findings and improve sensitivity, specificity and reproducibility of true findings. This session is focused on the methodological breakthrough that used multi-cohort/national Big Data collaborations to derive imaging and genetics signatures of complex illnesses from depression to cancer and translate them to guide personalized clinical decisions. The objective of our session is to encourage and disseminate novel analytical concepts, approaches, and applications to speed up the development of innovative technologies for hypothesis testing and data-driven discovery and translation to personalized medicine. Here we summarize the six submissions accepted for the session, with an emphasis on the diversity and coverage of the novel approaches. The accepted submissions were selected to cover novel analytic developments and applications with a focus on deriving novel risk measures for neuropsychiatric illnesses. The computational methods range from linear algebra to Artificial Intelligence and Machine Learning with imaging and omics data. *The first two contributions* focus on methodological developments intended to answer such fundamental questions as causality of identified genetic variants, preserving individual privacy in the Big Data genetic studies, and testing novel approaches for deriving genomic-trait association. *The second two contributions* report novel findings, including linking hypotheses generation and analyses across multiple Big Data samples. *The final two contributions* report on novel approaches for translating Big Data findings to the level of the individual in mental health and oncology.

2. Overview of Contributions

Childhood-to-adolescence is a critical period for brain development that corresponds to maturation of cerebral grey matter, that peaks at puberty, and maturation of cerebral white matter that peaks in late adolescence [1]. This supports the development and maturation of structural and functional networks that support higher cognitive skills [2-5]. It is also the period associated with development of lifelong, severe neuropsychiatric illnesses including autism spectrum disorder, schizophrenia, bipolar disorder, major depressive disorder and others [1, 6-11]. These illnesses are characterized by deviations from the normal brain maturation trajectory caused by action of risk factors that include genetic predisposition, pre/perinatal complications, childhood adversity and others. Early-life malnutrition has among the largest effect sizes and also is a key target for intervention and prevention. The manuscript by Gurkas and Karakurt describes a study where lifelong impact of early life malnutrition was quantified via EHRs data collected in adulthood. The greatest effects of childhood malnutrition in adults included problems with pregnancy/fetal abnormalities (20%), development of psychological/psychiatric illness (up to 16%), development of speech disorder (11%), followed by higher rates of various infection. Thus, childhood malnutrition can have lasting impact on both those who experienced it and their offspring.

The paper by Jacokes and colleagues considers advanced neuroimaging measures and blood-derived measures of gene expression to improve our understanding of autism spectrum disorder (ASD). Specifically, this paper uses logistic regression based on imaging and gene expression measures to predict ASD diagnosis, in a classification task, by using two different PCA-based approaches for feature reduction. The authors' integration of multiple methods is important for the field to advance. The lack of significant gene expression predictors suggests that brain microstructure anomalies may more tightly associated with ASD; even so, there may be a partial dissociation between blood-based and brain-based gene expression.

The paper by Noshin et al. explores the use of Electronic Health Records (EHR) to identify important diagnostic features for three types of Neuro-Degenerative Disorders (NDD), including Alzheimer's Disease (AD), Parkinson's Disease (PD), and other dementias (OD). By analyzing the EHR data from a cohort of 70,420 Alzheimer's Disease and Related Dementisa (ADRD) patients treated at Penn Medicine, the research aims to uncover key risk factors for these neurodegenerative disorders. The study employed both univariate and multivariate machine learning (ML) approaches and compared their performance in identifying risk features. A key finding is that the univariate approach was effective in uncovering rare but clinically important features specific to each disorder, while the features common across all methods represent the most robust indicators. The study also highlights the advantages and limitations of each ML method in the context of EHR data. This work is significant for researchers interested in using real-world clinical data to study neurodegenerative diseases, offering insights into the strengths and weaknesses of various ML approaches for ADRD and NDD research.

The effects of neuro-psychiatric illnesses on the brain are not regionally uniform. Neuropsychiatric disorders exert large pathological effects on some areas and circuits of the brain, while sparing others. Presently, Big Data meta-analytic studies of mental and neurological illnesses tabulate regional effect sizes using structural and/or functional brain atlases that are based on the anatomical boundaries, landmarks and connectivity patterns in healthy brains. Researchers have translated these findings to individual level predictors using approaches such as the Regional Vulnerability Index (RVI). RVI and other similar approaches quantify the agreement between individual brain patterns and the expected illness patterns identified by Big Data case control studies. Standard anatomical or connectomics-based atlases that were derived from healthy subjects are typically used to tabulate these effect sizes. However, these atlases are unlikely to capture the regional deficit pattern expressed in specific disorders, whereby the regions affected by illness may be averaged with regions that are spared, reducing the specificity and sensitivity of individual-level predictions. The study by Huang, Labate and colleagues posited that disorder-specific atlases derived using the Kullback-Leibler (KL) distance may offer a solution. KL-distance is a statistical measure of the dissimilarity between two arbitrary distributions. This offers a more stable approach to identifying areas of contrast between cases and controls than for example effect size-based measurements, because it is more stable in the presence of the non-Gaussian effects such as kurtosis, skewness and outliers. This study applied this approach to pilot a novel cortical template for Regional Homogeneity (ReHo) measurements in the subjects with the

Major Depressive Disorder (MDD). ReHo is measurement of homogeneity in the time course of blood oxygenation level dependence signal in functional MRI that was hypothesized to capture regional hypoperfusion deficits in this disorder. The MDD specific template the cerebral cortex was created by subdividing cortical landscape into contiguous region with 10 level. Each level constituted the compromise between the effects of MDD and size of the parcel to maximize contrast to noise ratio. They showed that the RVI metric--calculated using an MDD-specific parcellation--showed numerically higher effect sizes for separating patients and controls vs. those calculated using the standard Desikan-Killiany Atlas.

The contribution by He and colleagues addressed an important topic in the study of Alzheimer's disease (AD), which is to quantify Alzheimer's progression through multi-modal imaging-based pseudotime approaches. AD is a neurodegenerative disorder with no cures, and early detection is critical for successful intervention. This study explored pseudotime methods, which convert cross-sectional brain imaging data into 'faux' longitudinal data, to model the progression of AD and better understand how this complex process unfolds over time. Using data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort, the study evaluated pseudotime scores derived from individual imaging modalities and multi-modal data. The study found that most pseudotime analysis tools did not perform well on brain imaging data, with issues like reversed progression scores or poor distinction between diagnosis groups, likely due to assumptions designed for single-cell data. However, one tool showed promising results, where pseudotime from both single imaging modalities and multi-modal data captured the progression of diagnosis groups. Multi-modal pseudotime confirmed the hypothetical order of imaging phenotypes, and was primarily driven by amyloid and tau imaging, indicating their continuous changes across the full spectrum of Alzheimer's disease progression.

The manuscript by Ozdemir et al. tackles the long-standing question of predicting the future development of Alzheimer's disease (AD) in people who have mild cognitive impairment (MCI) – a condition that increases risk for AD, where people tend to develop AD at a rate of around 15% per year. The authors introduce a novel dynamic deep learning model for early prediction of AD (DyEPAD) to predict pro-gression from MCI to AD using EHR data. In the first step of DyEPAD, embeddings for each timestep or visit are captured through Graph Convolutional Networks (GCN) and aggregation functions. In the final step, DyEPAD employs tensor algebraic operations for frequency domain analysis of these embeddings, capturing the full scope of evolutionary patterns across all time steps. Their experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) and National Alzheimer's Coordinating Center (NACC) datasets show that their proposed model outperforms or is on a par with other state-of-the-art methods.

References

1. Kochunov, P. and L.E. Hong, *Neurodevelopmental and neurodegenerative models of schizophrenia: white matter at the center stage.* Schizophr Bull, 2014. **40**(4): p. 721-8.
2. Adibpour, P., et al., *Anatomo-functional correlates of auditory development in infancy.* Dev Cogn Neurosci, 2020. **42**: p. 100752.
3. Caffarra, S., et al., *Development of the visual white matter pathways mediates development of electrophysiological responses in visual cortex.* Hum Brain Mapp, 2021. **42**(17): p. 5785-5797.
4. Flechsig, P., *Developmental (myelogenetic) localisation of the cerebral cortex in the human.* Lancet, 1901. **158**: p. 1027-30.
5. Natu, V.S., et al., *Apparent thinning of human visual cortex during childhood is associated with myelination.* Proc Natl Acad Sci U S A, 2019. **116**(41): p. 20750-20759.
6. Rapoport, J.L., A. Addington, and S. Frangou, *The neurodevelopmental model of schizophrenia: what can very early onset cases tell us?* Curr Psychiatry Rep, 2005. **7**(2): p. 81-2.
7. Casey, B.J., J.T. Nigg, and S. Durston, *New potential leads in the biology and treatment of attention deficit-hyperactivity disorder.* Curr Opin Neurol, 2007. **20**(2): p. 119-24.
8. Kalia, M., *Brain development: anatomy, connectivity, adaptive plasticity, and toxicity.* Metabolism, 2008. **57 Suppl 2**: p. S2-5.
9. Feinberg, I., *Schizophrenia: caused by a fault in programmed synaptic elimination during adolescence?* J Psychiatr Res, 1982. **17**(4): p. 319-34.
10. Kochunov, P., et al., *Translating ENIGMA schizophrenia findings using the regional vulnerability index: Association with cognition, symptoms, and disease trajectory.* Hum Brain Mapp, 2020.
11. Kochunov, P., et al., *Ancestral, Pregnancy, and Negative Early Life Risks Shape Children's Brain Dis/Similarity to Schizophrenia.* Biol Psychiatry, 2023. **94**(4): p. 332-340.