Big Data Imaging Genomics

Peter Kochunov
Maryland Psychiatric Research Center, Department of Psychiatry,
University of Maryland School of Medicine, Baltimore, MD, USA
Email: pkochunov@som.umaryland.edu

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

John Darrell van Horn
Department of Psychology and School of Data Science,
University of Virginia, Charlottesville, VA, USA
Email: jdv7g@virginia.edu

Paul M. Thompson
Imaging Genetics Center, Mark and Mary Stevens Neuroimaging and Informatics Institute,
Keck School of Medicine, University of Southern California, Marina del Rey, CA, USA
Email: pthomp@usc.edu

This PSB 2022 session addresses challenges and solutions in translating Big Data Imaging Genomics research towards personalized medicine and guiding individual clinical decisions. We will focus on Big Data analyses, pattern recognition, machine learning and AI, electronic health records, guiding diagnostic and treatment decisions and reports of state-of-the-art findings from large and diverse imaging, genomics, and other biomedical datasets.

1. Introduction

Big Data studies in human health and disease have overcome formidable challenges to discover stable and reproducible phenotypic and genetic signatures of complex human diseases. This progress is being spurred by the development of imaging genetics approaches that integrate high-throughput imaging and multi-omics data (such as DNA sequences including SNPs and rare variants, RNA expression, methylation, epigenetic markers, proteomics, and metabolomics) and help make personalized medical decisions informed by Big Data knowledge. Historically, imaging genetics findings in complex illnesses have suffered from a substantial variability both within and across illnesses and the sources of heterogeneity have remained elusive. The presentations in this session demonstrate how Big Data collaborations such as the UK Biobank (UKBB), Enhancing Neuro Imaging Genetics through Meta Analysis (ENIGMA), Human Connectome Project (HCP), Alzheimer's Disease Neuroimaging Initiative (ADNI), Psychiatric Genomics Consortium (PGC), The Cancer Imaging Archive (TCIA) and others have enabled novel principled approaches to reduce false positive findings and improve the sensitivity, specificity and reproducibility of true findings. This session focuses on the methodological breakthroughs that have used multi-cohort multi-national Big Data collaborations to derive imaging and genetics signatures of complex illnesses from depression to cancer, and have begun to translate them to guide personalized clinical decisions.

The objective of our session is to encourage and disseminate novel analytic 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 seven submissions accepted for the session, with an emphasis on the diversity and breadth of the novel approaches. The accepted submissions were selected to cover novel analytic developments and applications using imaging and genetics data in complex disorders from Alzheimer's disease (AD) to major depressive disorder to cancer. The computational methods range from linear algebra to artificial intelligence and machine learning with imaging and omics data from the ADNI, PGC, ENIGMA, UKBB and TCIA. *The first three contributions* focus on methodological developments intended to answer such fundamental questions as causality of identified genetic variants, preserving individual privacy in Big Data genetic studies, and testing novel approaches for deriving genomic-trait associations. *The second two contributions* report novel findings, including linking hypothesis generation and analysis across multiple Big Data samples. *The final two contributions* report on novel approaches to translate Big Data findings to the level of the individual, in mental health and oncology.

2. Overview of Contributions

Genome-wide screening studies probe an association between the variance in a trait and individual polymorphic variation and may report ~10^{2-4} polymorphic variants that are significantly associated with a trait on the whole-genome level. A question that often remains unanswered is "*Are these variants causative*? This question can be answered directly by performing genetic manipulation of living organisms such as cells or animal models and noting the changes in the pathways associated with the expression of the trait. This, however, becomes impractical to perform for every association. It becomes even less practical in the multivariate imaging genetics case, where variance in 10^5 voxel-wise traits is being screened against 10^{6-7} polymorphic variants. The submission by Mo et al. [1] proposes to take advantage of the multi-omics nature of the Big Data imaging datasets to improve the specificity of these findings by identifying and eliminating *non-causative* association. Mo and colleagues propose a hypothesis driven test to screen for true causal variants. They propose use genetic variants as instrumental variables to test a biologically informed causative model where imaging data is used as an intermediate phenotype that acts between genes and biological traits such as cognition. They take advantage of the large sample size of Big Data studies to perform Mendelian Randomization to disturb the causative model and identify the genotypes that remain in causative association with biological traits by acting on intermediate phenotypes from those that are identified by chance. This highly novel and efficient approach can greatly reduce the number of genotypes of interest for further evaluation using biological models.

Big Data genetic analyses assemble very large sets of genotypes that are unique to an individual - unless they have an identical twin. The submission from Yamamoto et al. [2] aims to address a question of privacy in Big Data genomic analyses. Individuals who participate in research studies expect to remain unidentified but the genomic data which is unique to the individual always bears the risk for re-identification. This issue becomes even more important when these analyses are moved from research studies where individual subjects have signed an informed consent and acknowledge the potential risks of identification to the realm of personalized medicine where patients have additional rights regarding their privacy. Yamomoto et al. developed a method to perform family-based association tests while maintaining individual privacy by never disclosing the full genotypes. The exact algorithm is approximately three orders of magnitude faster than existing methods, and the approximation algorithm is ten times more efficient than the exact algorithm, while maintaining excellent accuracy.

The submission by Bao et al. [3] expands and exploits the granularity of the *morphometricity* approach for imaging genetics studies. The concept of *morphometricity*, inspired by the definition of heritability, is introduced to assess trait association with regional brain morphology. The information aggregated over the whole brain may carry the best signal but it lacks regional specificity, while regional variance may have more bearing on the disease but a higher level of noise may prevent the detection of a trait association. In this paper, the *morphometricity* is extended to a more local level based on regions of interest (ROI). A new method is proposed to identify a SNP-ROI association via regional *morphometricity* estimation of each studied single nucleotide polymorphism (SNP). An empirical study is performed, applying this method to the ADNI imaging genetics dataset. AD-related SNPs are shown to have higher overall regional *morphometricity* estimates than the SNPs not yet related to AD. This observation supports the value of imaging traits as targets in studying AD genetics.

In the second part of the presentations, we explore *two submissions* that report novel findings in Big Data samples using state-of-the-art methods. Bao et al. [4] present new results, identifying highly heritable traits that are informative of amyloid accumulation. Their study aims to use heritability estimates to prioritize amyloid imaging quantitative traits (QTs) for subsequent imaging genetics interrogation. Regional imaging QTs are often computed using predefined brain parcellation schemes such as the Automated Anatomical Labeling (AAL) atlas. However, the power to dissect genetic underpinnings underlying QTs that are defined in such an unsupervised fashion could be negatively affected by heterogeneity within the regions in the partition. To overcome this limitation, a novel method is proposed to identify highly heritable QTs by extracting regions containing spatially connected voxels with high heritability. An empirical study of the ADNI amyloid imaging and whole genome sequencing data demonstrates that the regions defined by this method yield much higher estimated heritability than the AAL regions. The method yields powerful imaging QTs to gain new insights into the phenotypic characteristics and genetic mechanisms of the brain.

The submission by Nir at al. [5] analyzed the UK Biobank dataset to evaluate the effects of ApoE genotypes on subcortical brain iron load and microstructure. ApoE4 is known to increase risk for AD, while the rarely studied ApoE2 genotype is thought to be protective. Nir et al. hypothesized that a person's ApoE genotype might modulate risk for AD and other neurodegenerative disorders by disrupting iron homeostasis. The accumulation of iron in the cellular and extracellular spaces leads to high oxidative stress burden and may interfere with protein folding processes. They tested this hypothesis in a large (N = 27,535) and inclusive cohort of mainly healthy individuals ascertained by the UKBB. The UKBB advanced imaging protocol included susceptibility weighted imaging (SWI) which was used to create quantitative susceptibility maps (QSM). These provide a sensitive biomarker of iron accumulation in brain tissue. Small but significant effects of these genotypes were detected in QSM measures of the hippocampus and amygdala.

In the last part of the session, we present *two submissions* that describe how Big Data driven findings may be translated to the level of individual to bring Big Data power to personalized medicine decisions. Kochunov et al. [6] used the brain deficit patterns reported in major depressive disorder (MDD) by ENIGMA, and examined these patterns in the independent UKBB dataset. They separated UKBB participants based on the level of depressive symptoms experienced by the individual and showed that the agreement between brain patterns and the expected deficit patterns observed in MDD had a strong overlap. This finding suggests that subjects whose brain patterns were more closely aligned with those observed in MDD patients are more likely to experience depressive symptoms - even in the absence of clinical diagnosis. They further showed that the similarity index they derived – known as the ENIGMA Dot Product (EDP) – has higher sensitivity to symptom severity than a genetic risk score for MDD. Finally, they showed that the overlap between individual deficit patterns goes beyond depression: subjects whose brains more closely resembled the characteristic patterns found in MDD, performed more poorly on cognitive tests that are impacted in depression.

The last submission of this session [7] aims to aid in personalized oncology decision-making and improve cancer care by automatically identifying phenotypes that are shared across multiple tumor types regardless of the origin. Chao and Belanger demonstrate that events such as genome-doubling - that occur in nearly every type of cancer, and have significant prognostic value - can be automatically identified from digital histology slices. Their study used data for 17 cancer types provided by the TCIA; they showed that a meta-learning approach can be used for separate-but-joint learning of the common features across these cancer types that are indicative of whether the genome doubling occurred in the individual sample. The study presents many technical points such as resolving batch effects, and examines the resolution and image characteristics necessary for translating the knowledge derived from Big Data studies to an individual histologic section.

The lectures presented in this session illustrate several important examples of the large-scale data analytic approaches valuable for associating the effects of genes on brain structure, connectivity, and disease features. They each describe a level of computational sophistication which is much needed for integrated machine learning and modeling of Big Data from the brain between spatial and genomic frames of reference. As the data from genomics becomes richer and less expensive to collect - in concert with the ever-increasing availability of state-of-the-art neuroimaging - such methods are likely to become widely used tools to characterize neurological syndromes and conditions and their genetic origins.

References

[1] Mo C, Ye Z, Ke H, Lu T, Canida T, Liu S, Wu Q, Zhao Z, Ma Y, Hong LE, Kochunov P, Ma T, Chen S: A new Mendelian Randomization method to estimate causal effects of multivariate brain imaging exposures. Pacific Symposium on Biocomputing 2022, 27.

[2] Yamamoto A, Shibuya T: Efficient Differentially Private Methods for a Transmission Disequilibrium Test in Genome Wide Association Studies. Pacific Symposium on Biocomputing 2022, 27.

[3] Bao J, Wen Z, Kim M, Saykin AJ, Thompson PM, Zhao Y, Shen L, for the Alzheimer's Disease Neuroimaging Initiative: Identifying imaging genetic associations via regional morphometricity estimation. Pacific Symposium on Biocomputing 2022, 27.

[4] Bao J, Wen Z, Kim M, Zhao X, Lee BN, Jung S-H, Davatzikos C, Saykin AJ, Thompson PM, Kim D, Zhao Y, Shen L, for the Alzheimer's Disease Neuroimaging Initiative: Identifying highly heritable brain amyloid phenotypes through mining Alzheimer's imaging and sequencing biobank data. Pacific Symposium on Biocomputing 2022, 27.

[5] Nir TM, Zhu AH, Gari IB, Dixon D, Islam T, Villalon-Reina JE, Medland SE, Thompson PM, Jahanshad N: Effects of ApoE4 and ApoE2 genotypes on subcortical magnetic susceptibility and microstructure in 27,500 participants from the UK Biobank. Pacific Symposium on Biocomputing 2022, 27.

[6] Kochunov P, Ma Y, Hatch KS, Schmaal L, Jahanshad N, Thompson PM, Adhikari BM, Bruce H, Chiappelli J, Goldwaser EL, Sotiras A, Ma T, Chen S, Nichols TE, Hong LE: Separating Clinical and Subclinical Depression by Big Data Informed Structural Vulnerability Index and Its impact on Cognition: ENIGMA Dot Product. Pacific Symposium on Biocomputing 2022, 27.

[7] Chao S, Belanger D: Generalizing Few-Shot Classification of Whole-Genome Doubling Across Cancer Types. Pacific Symposium on Biocomputing 2022, 27.