ABSTRACTS

SESSION I: RECENT ADVANCES IN STATISTICAL METHODS FOR MENTAL HEALTH SERVICES RESEARCH

Organizer: Dr. Elizabeth Stuart, Johns Hopkins University

Melanie Wall, PhD, Columbia University
Title: Data Science as the Engine for a Learning Health Care Service System for First Episode Psychosis in Coordinated Specialty Care

Abstract
A key initiative in research focused on treatment for first episode psychosis (FEP) is improving the implementation of evidence-based coordinated specialty care (CSC). One area of improvement is expected to come from improved data analytics facilitated by linking different clinical sites through common data elements and a unified informatics approach for aggregating and analyzing patient-level data. Through an NIMH-funded network and partnerships with the New York State Office of Mental Health and the Columbia University Department of Psychiatry, data science is contributing to a learning health care model. A few examples will be presented, including to what extent predictive modeling of patient-level outcomes based on background variables collected at intake and throughout care can be used to differentiate individuals in a way that is useful. Presentation of results will focus on interpretability of differential prediction across sites and usefulness for facilitating service decisions.

Yuanjia Wang, PhD, Columbia University
Title: Machine Learning Approaches for Optimizing Treatment Strategies for Mental Disorders

Abstract
Among currently available pharmacological and behavioral interventions for mental disorders, no single therapy is universally effective. Moreover, treatment responses are far from adequate across mental disorders. As such, there is an urgent need to optimize treatment responses. Various factors appear to be associated with positive treatment responses, thus providing evidence for improving response rate by incorporating patient-specific characteristics in treatment decisions in an effort to achieve precision psychiatry. However, individualized treatment decision-making for mental disorders faces challenges of extensive diagnostic heterogeneity, substantial between-patient variation in biological and clinical disease manifestation, and mismatch between diagnostic categorization and the underlying pathophysiology. We propose novel machine learning methods to address emerging challenges through probabilistic generative models and neural networks. We discuss several studies to discover reliable individualized treatment strategies that factor in a patient's clinical, psychosocial, and biological markers, and integrate evidence from multidomain data sources and multiple studies to increase generalizability and reproducibility.

Munmun De Choudhury, PhD, Georgia Tech
Title: Employing Social Media to Improve Mental Health: Harnessing the Potentials and Avoiding the Pitfalls

Abstract
Social media data is being increasingly used to computationally learn about and infer the mental health states of individuals and populations. Despite being touted as a powerful means to shape interventions and impact mental health recovery, we understand little about the theoretical, domain, and psychometric validity of this novel information source, or its underlying biases when appropriated to augment conventionally gathered data, such as surveys and verbal self-reports. This talk presents a critical analytic perspective on the pitfalls of social media signals of mental health, especially when they are derived from "proxy" diagnostic indicators, often removed from the real-world context in which they are likely to be used. Then, to overcome these pitfalls, this talk presents results from two case studies, where computational algorithms to glean mental health insights from social media were developed in a context-sensitive and human-centered way, in collaboration with domain experts and stakeholders. The first of these case studies, a collaboration with a health provider, focuses on the individual perspective, and reveals the ability and implications of using social media data of consenting schizophrenia patients to forecast relapse and support clinical decision-making. Scaling up to populations, in collaboration with a federal organization and toward influencing public health policy, the second case study seeks to forecast nationwide rates of suicide fatalities using social media signals, in conjunction with health services data. The talk concludes with discussions of the path forward, emphasizing the need for a collaborative, multidisciplinary research agenda, while realizing the potential of social media data in health – one that incorporates methodological rigor, ethics, and accountability, all at once.

SESSION II: STATISTICAL METHODS FOR GENERATING RELIABLE AND REPRODUCIBLE FINDINGS FROM NEUROIMAGING DATA
Organizer: Dr. Ying Guo, Emory University

Bin Yu, PhD
University of California, Berkeley
Title: Veridical Data Science for Biomedical Research with an Application to DeepTune for Characterizing V4 Neurons

Abstract
"A.I. is like nuclear energy – both promising and dangerous" – Bill Gates, 2019

Data science is a pillar of A.I. and has driven most of recent cutting-edge discoveries in biomedical research. In practice, data science has a life cycle (DSLC) that includes problem formulation, data collection, data cleaning, modeling, result interpretation, and the drawing of conclusions. Human judgement calls are ubiquitous at every step of this process, e.g., in choosing data cleaning methods, predictive algorithms, and data perturbations. Such judgment calls are often responsible for the "dangers" of A.I. To maximally mitigate these dangers, we developed a framework based on three core principles: predictability, computability, and stability (PCS). Through a workflow and documentation (in R Markdown or Jupyter Notebook) that allows one to manage the whole DSLC, the PCS framework unifies, streamlines, and expands on the best practices of machine learning and statistics – bringing us a step forward toward veridical data science.

We will illustrate the PCS framework in the modeling stage through the development of DeepTune images for characterization of neurons in the difficult V4 area of the primary visual cortex.

Thomas Nichols, PhD
University of Oxford
Title: The Impact of Methodological Variation in fMRI Inferences

Abstract
Challenges around reproducibility often focus on whether independent researchers can collect new data and produce a result consistent with an existing finding. This talk will focus on an even more fundamental type of reproducibility, the consistency of results when using the very same data but with different methodological pipelines. We took publicly available data for three different published task fMRI studies and systematically reanalyzed the data with three commonly used software packages, AFNI, FSL, and SPM. We tried to follow the published analysis (each originally conducted in one of AFNI, FSL, and SPM), but also made minor adjustments to the three software's pipelines to render them as similar as possible. Comparing unthresholded and thresholded statistical maps, we found surprisingly large differences between the different packages. Further extending this work, we have modified the pipelines to use a common (fMRIprep) preprocessing and still found substantial differences. This talk will discuss the origins of the differences and potential approaches to mitigate and ultimately accept these differences. This represents joint work with Alex Bowring and Camille Maumet.

Martin Lindquist, PhD
John Hopkins University
Title: The Role of Statistics in Large Complex Neuroimaging Studies

Abstract
Neuroimaging datasets are growing increasingly large and complex. With this new reality comes the need for principled statistical methods and thinking to analyze the resulting data and identify meaningful effects. Some commonly used statistical methods will carry over to this new paradigm, while others will need refinement. In this talk, we will seek to highlight some of the opportunities and challenges that lay ahead. We will also touch upon issues related to reproducibility, reliability, and causality, as well as the importance of evaluating results across different settings.

SESSION III: STATISTICAL TESTING AND POWER ANALYSIS FOR HIGH-DIMENSIONAL NEUROIMAGING MENTAL HEALTH DATA
Organizer: Dr. Dulal K. Bhaumik, University of Illinois at Chicago

Dulal K Bhaumik, PhD
University of Illinois at Chicago
Title: Power Analysis for High-dimensional Neuroimaging Studies

Abstract
We consider the problem of sample size determination for a neuroimaging study involving multiple comparisons. Neuroimaging data arise naturally in the context of developing biomarkers while comparing a neurological disease group with healthy control in a randomized clinical trial. The standard methodology for power analysis cannot be used or extended in such cases, as multiple comparison involving high-dimensional data is more complex and requires much more parametric information, rather than simply providing the effect size that often plays the vital role in a traditional power analysis problem. This talk will introduce some concepts like varying effect size and null and non-null distributions, while exploring the complexity of relevant data. It systematically develops a sample-size determination approach that has a simple interpretation and satisfies some desirable intuitive properties, like a large-effect size requires a small sample; more sample provides better power. Results will be illustrated with a real data set.

Rajesh R Nandy, PhD
University of North Texas
Title: A Semi-parametric Approach to Solve the Multiple Comparison Problem in Analyzing Functional MRI Data

Abstract
In conventional stimulus-driven fMRI data analysis, voxels are declared to be active by setting a threshold for the relevant statistic based on a chosen level of significance. However, in fMRI data, there are three factors that pose challenges in setting the right threshold. First, the fMRI time series at an individual voxel has strong temporal autocorrelation that needs to be estimated. The second factor is the multiple comparisons problem arising from simultaneously testing tens of thousands of voxels for activation. A common way in the statistical literature to account for multiple testing is to consider the family-wise error rate (FWE), which is related to the distribution of the maximum observed value over all voxels. The third problem, which is not mentioned frequently in this context, is the effect of inherent low-frequency processes present even in resting-state data that may introduce a large number of false positives without proper adjustment. We present an efficient semi-parametric method using resampling of normalized spacings of order statistics to address all three problems mentioned above. The correction for temporal autocorrelation is not critical in this approach.

Deepak N Ayyala, PhD
Augusta University
Title: Adjusting for Confounders in Cross-correlation Analysis of Resting-State Networks

Abstract
Resting-state network (RSN) analysis investigates spontaneous brain activity when the brain is not subjected to any external stimuli. The interest in RSN analysis lies primarily in understanding the interaction between different brain regions that occur while the brain is at rest, i.e., not prompted by external tasks. Testing for functional consistency in RSN requires analysis of time series patterns for multiple time series signals emanating from the different brain regions. An approach for studying reproducibility is testing for stability in the cross-correlations function of the multiple time series signal. However, often the testing procedures do not adequately account for the temporal dependence in the signal and may lead to erroneous conclusion, particularly in the presence of confounders such as scan-to-scan and visit-to-visit variation. In this talk, we present a general paradigm for testing for such confounders in the cross-correlation analysis. Merit of the proposed model is demonstrated via simulation, and the proposed test is shown to have reasonable type I error and power under a variety of dependence structures for the multivariate signals. The methodology is then applied to the motivating data set involving a motor network. It is shown that unless properly controlled, confounders can significantly affect the test of reproducibility of the network. Once the analysis is adjusted for confounders, the findings reaffirm the conventional wisdom about reproducibility of RSN.

SESSION IV: RECENT STATISTICAL DEVELOPMENTS IN IMAGING GENETICS
Organizer: Dr. Wes Thompson, PhD, University of California San Diego (UCSD)

Armin Schwartzman, PhD
University of California San Diego (UCSD)
Title: Estimating the Fraction of Variance of Cognitive Traits Explained by High-dimensional Genetic and Neuroimaging Measures

Abstract
The fraction of variance explained (FVE) by a model is a measure of the total amount of information for an outcome contained in the predictor variables. It is a fundamental quantity in much of mental health-related research, particularly genome-wide association studies (GWAS) and neuroimaging studies. In both of these domains, the number of predictors is extremely large, in the order of thousands to millions, far larger than the number of subjects. As a result, the effects of specific loci are extremely difficult to identify. In contrast, this talk shows that the FVE can be reliably estimated from data, even if only univariate summary statistics are available. For this we use an estimator called GWAS heritability (GWASH), originally developed to estimate the SNP heritability in GWAS; that is, FVE among all genetic SNPs in aggregate, regardless of significance. In this talk, we further show how the GWASH estimator can be adapted to the neuroimaging setting in order to estimate the FVE of an outcome among all locations in anatomical or functional brain images. The estimated FVE in GWAS and brain imaging can provide important insights into the amount of information about a cognitive outcome that is encoded in both these types of high-dimensional data.

Chun Fan, PhD
University of California San Diego (UCSD)
Title: Efficient Vertexwise/Voxelwise Imaging GWAS for Large-scale Heterogeneous Population Imaging Data and Enabling Downstream Multivariate Inference

Abstract
Although performing genome-wide associations is the very first step to establish the relationships between genetic variants and imaging measurements, the double curse of dimensionalities has significantly constrained how the imaging genome-wide association studies were conducted. Linear mixed-effects models that handle the relatedness between individuals, population structure, and repeated measures have been used sparsely. Meanwhile, the dimension reduction techniques were frequently used on the imaging measurements first in order to make the GWAS manageable. In this talk, we present our newly developed algorithm that can perform imaging GWAS across voxels and vertices fast and accurately. First, we estimate the variance components through methods of moments and refined with profile likelihood, assuming there are only finite sets of parameters across all voxels/vertices due to tissue properties. Second, we use GRAMMAR-Gamma approximation on the genetic variants, avoiding the need for inverting estimated variance in each genetic association. This new approach can enable researchers to obtain GWAS summary statistics to the resolution in the voxel level that can be further used for downstream multivariate inference, including loci discovery, trait prediction, and estimating genetic correlations.

Kevin Anderson, PhD
Harvard University
Title: A Platform for Imaging Genetic Study of Brain Aging

Abstract
Late life is marked by a decline in brain health. Loss of gray and white matter, neurovascular lesions, and reduced neurotransmitter tone are all hallmarks of advanced aging that are associated with decline in cognitive abilities. Recent large-scale data collections provide a new opportunity to understand brain aging across biological levels of analysis, from differences in gene expression by age to patterns of brain anatomy and function. This talk will present a platform for the study of aging that integrates imaging genetic data from the UK Biobank (N=40,000-500,000) and post-mortem measures of gene expression in the human brain (N>2,000). The platform enables interactive exploration of the “by age” component for thousands of neuroimaging, health, and behavioral measures, with the ability to visualize and analyze interactions with genetic and environmental variables. Brain transcriptomic analyses combine multiple post-mortem RNAseq datasets, including PsychENCODE, BrainSeq, NIH GTEx, and ROSMAP collections, which aim to facilitate deeper insight into specific gene and biological processes that vary with age. We will highlight through example analyses how dynamic and interactive data visualization can increase the pace of scientific insight and reduce barriers between historically siloed disciplines and data.