1-4-2018

Mental Health Data Science Overview

"There is little doubt that data science, an interdisciplinary field that brings together many data analysis approaches to help make sense of massive amounts of data, could provide new tools for exploring normal and abnormal human behavior and brain function." *JAMA Psychiatry* (2016) Why Psychiatry Needs Data Science and Data Science Needs Psychiatry by Torous and Baker.

Making sense and use of all types of data related to mental health research questions and clinical practice is what biostatisticians and data managers working in the area of psychiatry have been doing successfully for many years.   Biostatistics and Data Management have long been core areas of expertise, innovation, and collaboration within the Department of Psychiatry at Columbia University and the New York State Psychiatric Institute. So, why do we need a new name, Mental Health Data Science? As technologies for real-time and dynamic data collection have exponentially expanded over the last decade, the complexity and sheer size of this "Big Data" has expanded as well. For example, electronic medical records, multimodal brain imaging, ecological momentary assessment via smart phones, and neurological monitoring devices all pose serious collection, processing, and analytic challenges that must be addressed if the field of psychiatry is to make optimal use of the information encoded in these sources. In response, the biostatistical toolbox used in psychiatry has expanded to include high-dimensional machine learning methods, such as supervised and unsupervised prediction modeling, for making sense of this expanding array of data, translating it into useful answers to psychiatric research questions, and ultimately impacting clinical and policy decisions.

Our new name, Mental Health Data Science, exemplifies our mission to embrace new and emerging data collection and processing technologies--specifically with applications in mental health--and reflects our expertise in providing and developing analytic tools for making sense of these complex data domains while remaining grounded in the traditional, trusted, and highly effective core biostatistical tools that we have developed and implemented for years. Furthermore, we continue to focus on and implement careful archiving and documentation of computer code for all steps in processing and data analysis in order to ensure reproducibility of statistical results and efficient collaboration, for all research studies.

Mental Health Data Science is organized as two groups: Biostatistics lead by Dr. Melanie Wall, and Data Management lead by Dr. Howard Andrews.

The **Biostatistics group** includes 13 full time biostatisticians with expertise in statistical learning, analytics for personalized medicine, prediction modeling, network analysis, novel design and analysis of clinical trials, psychometrics, longitudinal data analysis, mediation causal analysis, propensity score analysis, survival analysis, functional data analysis, independent component analysis of imaging data, latent variable modeling including factor analysis and latent class analysis, structural equation modeling.

The **Data Management Group** (DMG) works closely with investigators and statisticians to design and develop secure, web-based systems for research data capture and management, project coordination, report generation, and the production of self-documenting data files for statistical analysis. Using advanced data systems and programming, the DMG utilizes decades of experience to maximize efficiency, minimize cost and facilitate the timely delivery of high quality, analytic-ready data files. The DMG builds data and project management systems for clinical trials, cross sectional and longitudinal studies, and is experienced in the processing and secondary analysis of large public and health-related data sets.