**niCHART: From Neuroimaging Big Data to Individualized Imaging Signatures of Disease**

*Guray Erus1, Ahmed Abdulkadir1, Gyujoon Hwang1, Zhijian Yang1, Mathilde Antoniades1, Junhao Wen1, Ioanna Skampardani1, Daniel Wolf2, Yong Fan1, Haochang Shou3, Theodore Satterthwaite4, Ilya Nasrallah1,5, Russell Shinohara3, Christos Davatzikos1*

*1Artificial Intelligence in Biomedical Imaging Laboratory (AIBIL), Center for Biomedical Image Computing and Analytics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, USA*

*2PennAlzheimer’sDiseaseResearchCenter,UniversityofPennsylvania,Philadelphia,PA,USA*

*3Department of Statistics and Data Science, University of Pennsylvania, Philadelphia, PA 19104, USA*

*4Department of Psychiatry, University of Pennsylvania, Philadelphia, PA*

*5Department of Radiology, University of Pennsylvania, Philadelphia, USA*

**Introduction**

Neuroimaging has been widely adopted by studies of brain development and aging, as well as of neurologic, neuropsychiatric, and neurodegenerative diseases. This rapidly growing web of data acquired over decades in multiple studies, in conjunction with the advent of advanced machine learning methods, has created tremendous potential for knowledge discovery at a large scale. Machine learning (ML), in particular, has offered great promise for precision diagnostics as well as derivation of imaging signatures of diseases, disease subtypes, endophenotypes, and predictive models. We describe the niCHART (NeuroImaging Computational Harmonization and ARtificial intelligence Toolbox), an ecosystem of algorithms and software implementations for data pooling, statistical harmonization, and ML-based analyses across studies. After integration of distinct existing components, niCHART will enable large-scale analyses of multi-modal brain MRI data, and mapping of such data into a dimensional system of neuroimaging signatures implemented by our ML models.

**Methods**

The axes of niCHART dimensional system represent two types of information: 1) a panel of structural (sMRI), diffusion (dMRI) and functional (fMRI) imaging derived phenotypes (IDPs), and 2) complex ML-based imaging signatures (ML-IDPs) derived from carefully processed and curated data (Fig. 1). The image processing module integrates advanced, validated pipelines [1-3], as well as automated tools for quality control [4], that have been publicly released and applied to large scale studies. A statistical harmonization method for multi-site MRI data adapted for nonlinear age trends is utilized for explicit removal of site effects [5]. ML-IDP models [6-9] map the high dimensional image data into the niCHART reference frame [10]. Dimensions of niCHART capture and quantify multi-variate imaging patterns of brain change reflecting the heterogeneity of brain aging, neurodegeneration, as well as neuropsychiatic disorders.

**Results**

niCHART reference data consist of pooled and harmonized imaging data of 65,693 individuals (84,211 time points) from 22 studies, with sMRI for all time points, and dMRI and fMRI for smaller subsets. Derived imaging variables include data-driven and atlas based multi-scale brain parcellations, structural covariance units, connectivity measures, and brain networks. The dataset is used to establish normative ranges of brain structure and function, as well as a rich set of ML-IDPs (Fig 2). In particular, we constructed two types of ML-IDPs. Supervised ML-IDPs, indices derived based on labels of interest, e.g. disease or clinical progression status, were applied to train models for brain aging throughout the lifespan, schizophrenia, depression, autism spectrum disorder, and Alzheimer’s disease AD [6,7,8]. ML-IDPs using semi-supervised pattern clustering methods, which aim to uncover disease heterogeneity in a data-driven fashion, identified two distinct signatures of schizophrenia and their clinical correlates, and dissected the neuroanatomical heterogeneity of AD into 4 subtypes, with two distinct longitudinal pathways in MCI to AD progression [10]. The panel of reference ML-IDP values and pre-trained models allow users to place their images into the chart by computing the extent an individual expresses these imaging signatures.

**Conclusions**

niCHART offers a software suite, along with associated machine learning models, for analysis of neuroimaging datasets including sMRI, dMRI, and fMRI. Extendible ML models provide users with the opportunity to augment current imaging-derived phenotypes with a panel of ML-based imaging signatures associated with healthy and diseased brain states. Via statistical harmonization and machine learning, two central pillars in our system, niCHART has the potential to provide the machinery to derive biomarkers that focus on individual-centric analyses, and hence contribute to the concepts of precision medicine and dimensional phenomics.

***References***

*[1] Doshi, J., et al., (2016), MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. Neuroimage, 127: p. 186-195.*

*[2] Esteban, O., et al., (2019), fMRIPrep: a robust preprocessing pipeline for functional MRI. Nat Methods, 16(1): p. 111-116. 30.*

*[3] Cieslak, M., et al. (2021), QSIPrep: an integrative platform for preprocessing and reconstructing diffusion MRI data. Nat Methods, 18(7): p. 775-778.*

*[4] Satterthwaite, T.D., et al., (2019) Motion artifact in studies of functional connectivity: Characteristics and mitigation strategies. Hum Brain Mapp, 40(7): p. 2033-2051.*

*[5] Pomponio, R., et al., (2020), Harmonization of large MRI datasets for the analysis of brain imaging patterns throughout the lifespan. Neuroimage, 208: p. 116450.*

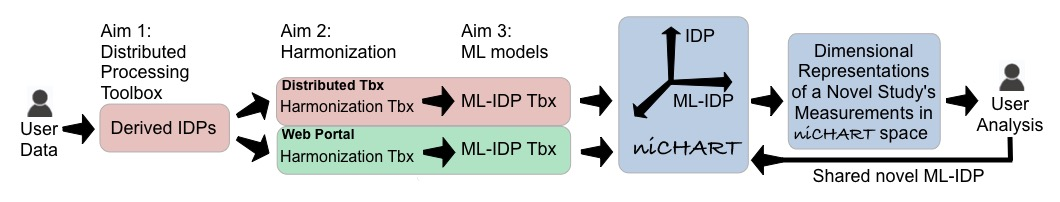
*[6] Rozycki, M., et al., (2018), Multisite Machine Learning Analysis Provides a Robust Structural Imaging Signature of Schizophrenia Detectable Across Diverse Patient Populations and Within Individuals. Schizophr Bull, 44(5): p. 1035-1044*

*[7] Davatzikos, C., et al., (2009), Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain, 132(Pt 8): p. 2026-35. 53.*

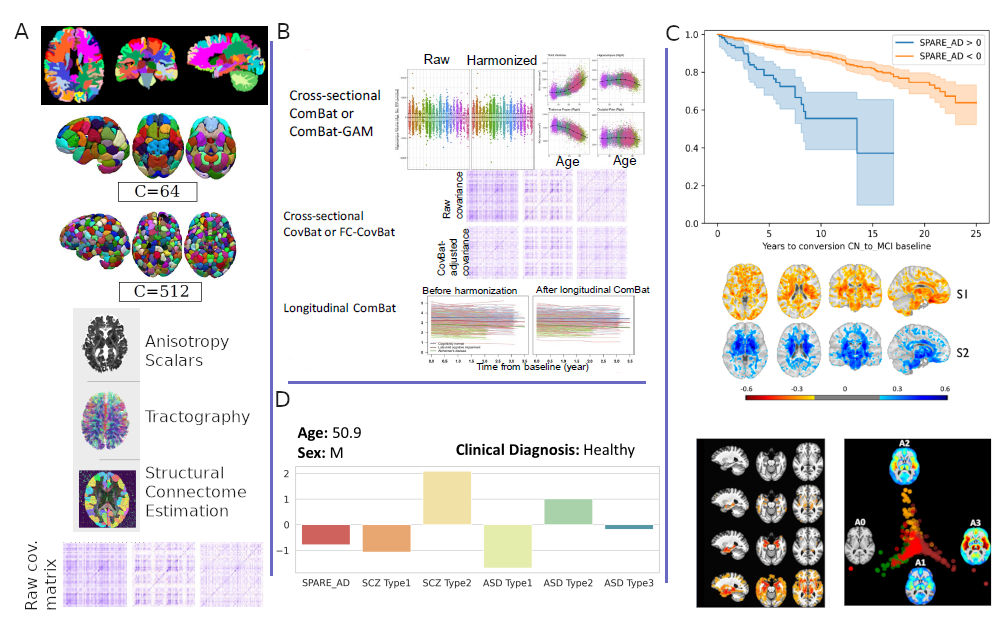
*[8] Yang, Z., et al. (2021), A deep learning framework identifies dimensional representations of Alzheimer’s Disease from brain structure. Nat Commun 12, 7065*

*[9] Hwang et. al. (2022), Three Neuroanatomical Endophenotypes of Autism Spectrum Disorder and Their Partially Overlapping Characteristics with Schizophrenia. Jama Psychiatry.*

*[10] Habes M. et al. (2021), The Brain Chart of Aging: Machine-learning analytics reveals links between brain aging, white matter disease, amyloid burden, and cognition in the iSTAGING consortium of 10,216 harmonized MR scans. Alzheimers Dement. ;17(1):89-102.*



***Figure 1:*** *Schematic of the image processing pipelines and software suite that allows researchers to use or derive, from pooled reference data and their data, a dynamically growing number of AI-based biomarkers of brain structure and function, which capture complex multi-variate patterns of typical and pathologic brain states. Dimensional niCHART serve as a canonical space within which researchers are able to harmonize their data with other studies using provided software tools, and apply a variety of ML models defined in that space.*



***Figure 2:*** *NiChart imaging phenotypes (A), statistical harmonization (B), ML-based imaging signatures (C), and an example of individualized ML-IDP panel (D). A (top to bottom): Structural regions of interests; data driven multi-scale non-negative matrix factorization components (the figure shows two levels of the hierarchical decomposition for 64 and 512 components); DTI image variables; functional connectivity variables. B: Statistical harmonization using ComBat, CovBat and longitudinal ComBat. C (top to bottom): survival curves of positive and negative SPARE-AD groups for conversion from MCI to AD; imaging patterns, identified using the semi-supervised pattern clustering method HYDRA, that differentiate the 2 Schizophrenia (SCZ) sub-types from reference CN subjects; imaging patterns, identified using the semi-supervised pattern clustering method SmileGAN, that differentiate the 4 AD sub-types from reference CN subjects; C4. Imaging patterns, identified using SmileGAN, that differentiate the 3 brain aging sub-types (A1-A3) from reference CN subjects (A0). D: Individualized ML-IDP panel for a single subject provided here for illustration purposes.*