PGC Secondary Analysis Proposal *(v2, revised 02-2015, pfs)*

Date	5/5/20
Title	ENIGMA PTSD and Bipolar cross-disorder vertex-wise machine learning project

Investigative Team. Underline PGC PI taking responsibility for all aspects of this proposal.

Name	**Email**	**Group**
Ling-Li Zeng	lingl.zeng@gmail.com; zengphd@nudt.edu.cn	ENIGMA Central + Bipolar
Chris Ching	cching78@usc.edu	ENIGMA Central + Bipolar
Paul Thompson	pthomp@usc.edu	ENIGMA Central + Bipolar
Xin Wang	xin.wang2@utoledo.edu	ENIGMA PTSD
Kevin Xu	Kevin.Xu@utoledo.edu	ENIGMA PTSD
Brian O'Leary	Brian.O'Leary@utoledo.edu	ENIGMA PTSD
Rajendra Morey	rajendra.morey@duke.edu	ENIGMA PTSD

Data access requested. No permission required for published results (only pre-publication)

Group	**Individual genotypes**	**Summary results**	**Permission from group?**	**Version (e.g., MDD2, SCZ3)**
ENIGMA PTSD	NA			
ENIGMA BD	NA			

A. Research Question, Goal, or Specific Aim

Provide a brief description (e.g., 1 paragraph) describing the aims of the proposal and the research questions to be addressed.

We propose to investigate brain shape/volume patterns (vertex-wise radial distance and log of Jacobian determinant for subcortical structures, and FSL/FreeSurfer vertex-wise cortical thickness, surface area, curvature, sulcal depth for cerebral cortex, as well as ROI-based gross volumes for subcortical structures and ROI-based cortical thickness and surface area) in participants with PTSD, BPD and controls using advanced machine learning/deep learning techniques.

The first question is whether brain shape/volume metrics can provide helpful information for the clinical diagnostic classification across disorders (PTSD, and BPD). Key to this analysis is determining whether brain shape/volume patterns are significantly different across PTSD, BPD, and healthy controls, and whether such patterns can successfully classify the participants with PTSD, BPD, and healthy controls by using supervised machine learning approaches (detailed in section below).

Further questions include whether cortical/subcortical shape/volume metrics can provide helpful information for delineating disorder subtypes. To address this issue, we will test whether the shape/volume metrics can cluster the participants with neuropsychiatric disorders into different subtypes by using unsupervised machine learning approaches. Based on the findings from the clustering analysis, we propose to investigate how data-driven clusters may be associated with clinical variables, such as medication or medication-free, first episode or recurrence, depression symptoms, and other available symptoms and clinical metrics available across both working group samples.

B. Analytic Plan

Provide a brief description of the analyses to be performed to address the research questions described above. Include relevant details e.g. phenotype definition, QC, analysis, plans to address population stratification and other confounders, power.

The phenotypes include available cortical and subcortical vertex-wise/ROI-based measures previously derived as part of previous ENIGMA PTSD and BD projects (specific measures detailed below). Both Working Groups have derived cortical measures and the BD Working Group has already generated subcortical shape attributes. These data will have already been quality inspected using standard ENIGMA pipelines across working groups.

The first question is whether the shape/volume metrics can provide helpful information for the clinical diagnostic classification across disorders. In the tests to address this issue, the challenges mainly come from small cross-disorder differences, divergence across multiple sites, and incompleteness of shape/volume metrics for individual subjects (some structures may be dropped for a given subject due to quality control). Thus, we should design/select appropriate machine learning methods to address these issues.

For example, we plan to 1) test a two-phase multi-view classification framework, in which a sub-classifier would be trained for each structure first, and then ensemble learning can be used at a decision level and can handle the incomplete data well, 2) implement an end-to-end deep learning framework for the shape data which projects 3D vertex-wise mesh data onto 2D planar images and then use 2D CNN for the classification. In this deep learning framework, low quality data could be considered as noise to enhance the generalizability of the classification models.

The second question is to determine the extent to which shape/volume metrics cluster subjects across disorders. In the clustering analysis, due to high dimensionality and high similarity between individuals of the shape/volume metrics, Laplacian Eigenmaps will be used for dimensionality reduction, which creates a representation for data lying on a low-dimensional manifold embedded in a high-dimensional space. Due to low Signal-to-Noise Ratio (SNR) of the shape data, many data points (individual subjects) may be considered as outliers in the clustering analysis. To automatically spot and exclude outliers from the analysis, we would implement and test density-based clustering algorithms, which can deal with cluster cores (high quality sample) and halos (noise sample). When the subjects were clustered into several groups on the basis of brain shape/volume patterns, we could examine the potential relationship of the clusters to clinical metrics such as medication, first episode/chronic, depression symptoms, etc.

C. Analytic Personnel

Indicate who will be responsible for performing the analyses.

Ling-Li Zeng, Christopher Ching, Paul Thompson (USC main analysis site), in consultation with all those listed above in the investigative team

D. Resources Needed

Describe the resources needed to achieve the aims of the analysis, including variables needed, analytic support, and any other issues that may affect the feasibility of the plan.

Individual subject subcortical and cortical ROI and vertex-wise shape metrics (such as vertex-wise radial distance and log of Jacobian determinant for subcortical structures, and together with FreeSurfer vertex-wise cortical thickness, surface area, curvature, sulcal depth for cerebral cortex), demographic and clinical variables (Covariates.csv) as previously prepared for the ENIGMA-PTSD cortical/subcortical papers (including gender, age, education, subtype, duration of illness, treatment, history of psychosis and other clinical measures when available). The image file list is as the following:

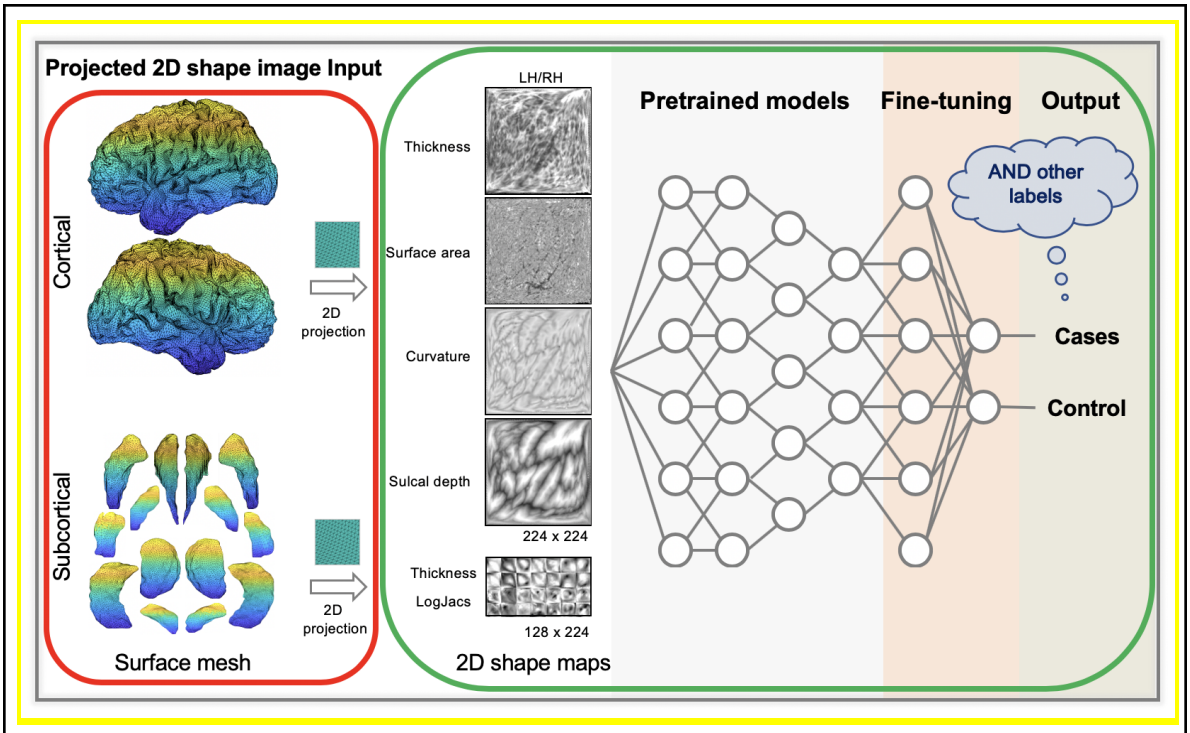
>Subcortical (when available):
ROI-based Gross Volumes (aseg.stats)
LogJacs_ID.raw
Thick_ID.raw
Curve_ID.ucf
Resliced_mesh_ID.byu
Resliced_mesh_ID.m
ID=10,11,12,13,17,18,26,49,50,51,52,53,54,58

>Cortical:
ROI-based thickness and surface area measurements (lh.aparc.stats; rh.aparc.stats)
/Surf/**.thickness
/Surf/**.curv
/Surf/**.sulc
/Surf/**.area
/Surf/**.orig
/Surf/**.pial
/Surf/**.sphere.reg
**=lh or rh

Individual subject shape data are essential to answer the research question. In accordance to previous ENIGMA PTSD and BD projects, data is anonymized, and the individual subject shape data would be used only for the purposes outlined here.

The individual data will reside on a research dedicated cluster in Dr. Paul Thompson's lab at the Imaging Genetics Center of the USC, which is password protected and only accessible by those directly involved in the analyses for this project. We will perform optimal mass transport on the secure USC cluster to generate 2D shape metric images for the deep learning input. The 2D images will be transferred to the **supercomputer** ***cluster in Hunan Center, China, which is password protected and only accessible by those directly involved in the analyses for this project.*** **The data processing steps are outlined in the below schematic with Red outlines indicating primary data storage and processing at USC, and green outlined steps to be carried out on the anonymized 2D images by Dr. LingLie Zeng.**

It should be noted that all individual 3D brain surface meshes are registered and downsampled to the FreeSurfer fsaverage6 template first, and then are projected to 2D planar meshes using optimal mass transport. Moreover, interpolation operation and scale standardization is conducted with the shape metrics (fsaverage6 space) to generate the final 2D RGB images. So it is impossible to reconstruct an individual brain surface mesh from the final 2D images, providing an additional layer of data protection when sharing such images.



F. Timeline

Estimated time required to complete the plan and write a paper (should be ≤ 6 months).

We hope to have the project reviewed and finalized by the group by the end of July, 2020. Initial data collection to start in June and continue for 6 months. We will set a data freeze date, beyond which no new data would be accepted. This is necessary to prevent delays and re-analyses. We will proceed with analyses after the data freeze day and plan to submit the manuscript for peer review in the first half of 2021.

F. Collaboration

The following is the standard PGC policy about secondary analyses. Any deviation from this policy needs to be described and justified, and could negatively impact the proposal PGC investigators who are not named on this proposal but who wish to substantially contribute to the analysis and manuscript may contact the proposing group to discuss joining the proposal.

G. Authorship

This is an extremely important part of this proposal. Describe how authorship will be handled in the manuscript resulting from this analysis. To avoid a revision, first review the authorship policy of the group(s) whose data you wish to analyze. Points to consider:

(a) are you following the authorship policies of the groups involved?

Yes

(b) will there be a writing group and if so, who will be included?

Yes, all those listed in the investigative team

(c) what groups or individuals will be listed as authors?

All those ENIGMA bipolar and PTSD working group members contributing data to the analysis and the investigative team

(d) will PGC members not listed as named authors be listed at the end of the manuscript?

Yes under the banner "for the ENIGMA Bipolar and PTSD Working Groups"

(e) will PGC members or groups be listed as "collaborators" on the PubMed abstract page?

(f) how will funding sources be handled or acknowledged?

All sources will be named in the acknowledgement section of any abstract, conference presentation or manuscript