Machine learning improved classification of psychoses using clinical and biological stratification: Update from the bipolar-schizophrenia network for intermediate phenotypes (B-SNIP)

Suraj Sarvode Mothi a,b,1, Mukund Sudarshan c,1, Neeraj Tandon a,d, Carol Tamminga e, Godfrey Pearson f,g, John Sweeney h, Brett Clementz h,i, Matcheri S. Keshavan a,*

a Department of Psychiatry, Beth Israel Deaconess Medical Center and Harvard Medical School, 75 Fenwood Rd, Boston, MA 02115, USA
b Baylor College of Medicine, Houston, TX, USA
c Psychiatry, UT Southwestern, Dallas, TX, USA
d Department of Psychiatry, University of Cincinnati, Cincinnati, OH, USA
e Department of Psychiatry, Yale University, New Haven, CT, USA
f Department of Psychiatry, University of Cincinnati, Cincinnati, OH, USA
g Olin Neuropsychiatry Research Center, Hartford, CT, USA
h Department of Psychology, Bio-Imaging Research Center, University of Georgia, Athens, GA, USA

1 SSM and MS are co-first authors.

A R T I C L E   I N F O

Article history:
Received 25 August 2017
Received in revised form 25 April 2018
Accepted 27 April 2018
Available online 26 May 2018

1. Introduction

Psychiatry continues to suffer from challenges to diagnostic validity due to lack of biological markers. Distinctions between diagnostic categories are still largely informed on symptom clusters, while underlying biology is complex and largely indiscernible. (Insel et al., 2010; Keshavan et al., 2013) While the reliability of diagnoses has improved with successive revisions of the classification systems such as the latest Diagnostic and Statistical Manual of Mental Disorders (DSM-V), whether these disorders have natural biological distinctions is uncertain. This is particularly true of psychotic disorders, which are currently classified into schizophrenia, schizoaffective disorder, and psychotic bipolar disorder following a century-old distinction by Moskowitz and Heim (2011); between disorders with predominantly thinking/cognitive versus affective symptomatology. However, these disorders have substantial overlaps in regard to etiology, biomarkers, treatment response, and outcome. As such, the current diagnostic scheme has sub-optimal predictive value in clinical settings (Tamminga et al., 2013). Psychiatry needs testable theories with higher levels of specificity. As the science of psychiatry matures, the field must become more self-critical about the validity of its theories (Kendler and Schaffner, 2011). In the light of modern neuroscience, psychiatry should encourage "proof of concept" approaches that focus on putting the "brains" back in psychiatry. Current classifications based on symptomatology alone, do not separate biologically distinguishable phenotypes. It would be important to know what combinations of biological factors in interplay with clinical manifestations might be useful in predicting severity of illness. Our proposed approach examines the precision with which "natural" boundaries of the illness maybe demarcated by a combination of such clinical and biological factors. Machine learning could be a useful tool in uncovering general principles without explicit instructions and allowing the data to "speak for itself" (Bzdok and Meyer-Lindenberg, 2017).

Several efforts have been made to propose a purely neuroscience-based classifications for psychotic disorders. (Orrù et al., 2012; Sun et al., 2015) Clementz, et al. recently conducted a k-means clustering analysis on the proband data used in the current analysis using cognitive and electrophysiological biomarkers and identified three "biotypes" that were largely orthogonal to the DSM diagnoses (Clementz et al., 2016). These biotypes showed significant differences in external validating measures such as brain morphology (Clementz et al., 2016) and resting-state fMRI (Meda et al., 2016). The biotype distinctions, however, were based entirely agnostic to clinical symptomatology. In practice, one needs the ability to distinguish between diagnostic categories by integrating both clinical and laboratory-based data to construct well-informed decisions. An integrated approach can facilitate the ability to individualize care, determine the course of illness and improve treatment response.

Machine learning (ML) offers much promise with the ability to take complex multivariate information for generating the "best-case" discrimination between diagnostic groups. It performs well when dealing with high dimensionality in features (Bengio et al., 2013) and also provides the ability to account for a large number of...
non-linear interactions between variables (Schölkopf et al., 1998). Such advantages offer the ability to integrate heterogeneous data sources and delineate non-linear relationships between multiple variables thereby striving towards more discernible distinctions between psychosis syndromes.

To the best of our knowledge, no prior study has examined the utility of ML across the psychosis spectrum using an exhaustive set of phenotypic constructs. In most cases, experiments involving ML, especially supervised techniques, aim to simply differentiate between psychosis and healthy controls (Iwabuchi et al., 2013; Orrù et al., 2012; Yang et al., 2010) thereby displaying good predictive power for a relatively simpler task. With the absence of a biological “ground truth,” supervised experiments are fundamentally limited by the quality of the diagnosis labels (Wolters et al., 2015). However, clinically, it is also a very important task to differentiate between different forms of psychosis.

Furthermore, many ML studies in psychosis rely on only one classification algorithm instead of trying multiple approaches, with only a few studies systematically comparing two or more algorithms (Castellani et al., 2012; Rathi et al., 2010). ML algorithms tend to overfit when the number of samples is too small or when not enough data is withheld to validate the model (Schnack and Kahn, 2016). A study with a larger sample would be less likely to produce a model that over-fits the data (Button et al., 2013; Varoquaux, 2017).

The fundamental goal of the present study is to delineate between psychosis subgroups using a “stratified” approach, by testing the performance of an unsupervised learning algorithm when feeding strata (clinical, EEG, cognition) of features in stages on a large sample of psychotic probands involving Schizophrenia (SZ), Schizoaffective (SZA) and Bipolar-Disorder (BP). The Bipolar-Schizophrenia Network for Intermediate Phenotypes (B-SNIP) consortium was formed to examine a broad array of biomarkers across psychotic disorders and to test the hypothesis that biomarker characteristics are homogeneous between phenologically derived DSM-IV psychosis diagnoses. We first used an unsupervised ML algorithm to identify different groups within the probands based on a range of modalities (laboratory bio-markers, DSM diagnosis labels, and clinical information); we then compared separability of these groups using a variety of external validators (structural MRI data, social functioning scales and smooth pursuit eye movements). Our aim was not to invalidate any one type of delineation (biotypes, DSM), but to determine if an approach that combines multiple families of features can better differentiate psychosis subgroups. We hypothesized that combining multiple sources of information about each patient will enhance differentiations. Applying a non-linear kernel allows the ML algorithm to linearize relationships and therefore account for complex interactions between variables (Hoffmann et al., 2008). Therefore, we also hypothesized that a kernelized algorithm will produce more distinct delineations.

2. Methods

2.1. Sample description

The study included 610 probands with a psychotic disorder: 251 with schizophrenia (SZ), 164 with schizoaffective (SZA) and 195 with bipolar disorder with psychosis (BDP) from the B-SNIP database on whom complete case of clinical measures and demographic information was available. A sub-sample of $n = 400$ of the probands had 3.0 Tesla structural MRI scans. 342 demographically comparable healthy controls (‘With no past/family history of psychotic disorders) were recruited. Recruitment details, inclusion/exclusion criteria are described in detail elsewhere. (Tamminga et al., 2013) Diagnosis was determined using the Structured Clinical Interview of the DSM-IV, which was reviewed by at least two experienced research clinicians to establish a consensus diagnosis.

2.2. Description of variables

2.2.1. Features used for clustering

2.2.1.1. Clinical measures. Positive and Negative Symptom Scale (PANSS), (Kay et al., 1987) Montgomery Asberg Depression Rating Scale (MADRS), (Montgomery and Asberg, 1979) and Young Mania Rating Scale (YMRS) (Young et al., 1978) were collected in all individuals with an Axis 1 clinical psychosis diagnosis and were administered by trained raters. Standardization of symptom ratings across sites was performed by periodic meetings for rater training, using established “gold standard” interviews. At study initiation, an in-person training session was held for all raters, with a requirement $>0.85$ reliability to be attained before deemed eligible to administer scales. Rater training was repeated annually to re-establish reliability. The resulting analytical database utilized for clustering, comprised of individual and composite items for all aforementioned scales.

2.2.1.2. Laboratory measures

2.2.1.2.1. Laboratory measures included Eye movement tasks. Prosaccades (3 blocks of 32 trials) assessing speed of visual orienting and an anti-saccades task (4 blocks of 20 overlap trials) assessing inhibitory control under perceptual conflict due to incompatibility between the orienting cue and the required response. For details of methods and BSNIP data with these tasks, see (Reilly et al., 2014).

2.2.1.2.2. Cognitive assessments. Stop signal task which measures an aspect of cognitive control, specifically the ability to suppress a prepotent behavioral response. For details on the paradigm and BSNIP data with this task, see Ethridge et al. (2014). Additionally the Brief Assessment of Cognition in Schizophrenia (BACS), a neuropsychological battery assessing multiple cognitive functions was administered. Test procedures and data from the BSNIP sample have been previously reported (Hill et al., 2013).

2.2.1.2.3. Auditory paired stimuli and oddball evoked brain responses. Evoked brain responses assess the neural dynamics of preparation for and recovery from auditory sensory activations. Methodology and data are reported elsewhere (Ethridge et al., 2015).

2.2.2. Validation measures

2.2.2.1. MRI structural Imaging. Subjects were scanned in 6 sites: Boston, Detroit, Baltimore, Hartford, Dallas, and Chicago. Sites used similar but slightly different MPRAGE acquisition parameters; full details are outlined previously (Mathew et al., 2014). The Alzheimer’s Disease Neuroimaging Initiative (ADNI) protocol was used at all sites to standardize imaging analysis (Jack et al., 2008). Boston was the neuro-image processing site, where images from all sites underwent a meticulous quality control process. All images were checked for scanner artifacts and were processed in auto-recon 1 in FreeSurfer v5.1 (Fischl, 2012). Trained and blinded raters, all reliable above 95%, edited images to remove any remaining non-brain tissue. After adequately cleaned for segmentation, images were then processed through auto-recon 2 and 3 to extract regional cortical thickness (CT) measurements. For consistency, all analysis mentioned in Section 2.5 using CT measurements as dependent measures are adjusted for age as a continuous covariate and sex, race and site differences treated as categorical “dummy-coded” variables.

2.2.2.2. Measures of social functioning. Deficits of social functioning were assessed using the Birchwood SFS (Birchwood et al., 1990) and the global assessment of function (GAF) as assessed over a composite scale of 0–100 in the DSM-IV-TR.

2.2.2.3. Smooth pursuit testing. Impaired sustained pursuit maintenance in eye tracking has proved a promising candidate as an intermediate phenotype in schizophrenia schizoaffective disorder and psychotic
bipolar disorder. (Clementz and Sweeney, 1990; Holzman et al., 1974; D. R. Rosenberg et al., 1997; Sweeney et al., 1994) Smooth pursuit tracking paradigms and data from the B-SNIP sample has been detailed previously (Lencer et al., 2015).

2.3. Preprocessing

An issue for biological data, especially neuroimaging, is that the amount of nuisance variation is usually higher than at least discernibly cohesive clinical symptom manifestations. Thereby, to avoid the clustering algorithm to be influenced by nuisance factors, we a priori decided to withhold structural MRI measures and smooth pursuit laboratory outcomes for validation purpose only. We also validated cluster stability of our groupings for clinical usefulness based on functioning clinical outcomes using SFS total and GAF scores.

We used the following datasets as inputs for the unsupervised machine learning analyses. A) Laboratory data: Dataset containing \( n = 39 \) biomarker measures comprising of EEG, pro- and anti-saccade measures, the stop signal task, and 6 cognition test scores from the BACS test (see Clementz et al., 2016 (Clementz et al., 2015) for a full description). B) Clinical dimensional data: Clinical ratings were used from YMRS (\( n = 11 \) items), PANSS (\( n = 7 \) positive items; \( n = 7 \) negative items; \( n = 16 \) general items), MADRS (\( n = 10 \)) and with a single variable containing each subject’s categorical DSM IV diagnosis, with three possible values for each subject. 1: Bipolar Disorder 2: Schizophrenia 3: Schizoaffective Disorder. C) Composite data: Clinical dimensional data + Laboratory data.

2.4. Unsupervised learning

To systematically evaluate clustering performance, we conducted our experiment incrementally, by feeding strata (Laboratory, clinical) of features independently, at first, ultimately leading to a “composite” clustering approach. Regardless of input strata, validation constructs (structural MRI, stop-signal and functioning measures) remained uniform and was evaluated for differences between clusters generated at every stage. For simplicity, we report here validation results for clusters produced through the agglomerative approach only. Using the Scikit-learn package in the Python programming language, unsupervised learning was applied with steps iterated in Fig. 1. The following three procedures were applied to each input strata:

2.4.1. Robust Scaler

As features with differing scales were utilized for clustering, all data were scaled using a robust scaler, assuring that each variable had approximately the same scale (\( \mu = 0, \sigma = 1 \)). The median values -10% and above 90% were excluded when computing the mean and sd for scaling. The resulting scale is robust to outliers which lie in the 10% tails, thereby not affecting the means or standard deviations.

2.4.2. Cluster construction

Spectral embedding was performed on binary and multinomial features in our dataset in order to embed them in a continuous space. These embedded features were concatenated with the continuously variable features. It is common practice for Principal Component Analysis (PCA) to be used to project high-dimensional datasets to a lower dimensional subspace prior to applying any clustering methods. (Ding and He, 2004; Jolliffe, 2014) Therefore, Kernelized PCA was applied using the Radial Basis Function (RBF) kernel, as it is the best-known kernel to demonstrate novelty detection in high dimensional feature space (Hoffmann, 2007; Mika et al., 1998). To capture enough variance in the data, we kept 10 principal components which was the smallest number of components required to retain 95% of the Frobenius norm. We then ran K-Means clustering, calculated the average silhouette scores (As mentioned in Section 2.4.3) across all points along with a variety of clustering performance metrics (Supplemental Table 1), and created silhouette plots for each of the k clusters. To facilitate a fair comparison between the different strata of features, we first established a baseline clustering using only the laboratory data (EEG, saccade, stop signal, and cognition); specifically, the data used to create the biotypes framework in Clementz et al. (2016) as we consider it to be a valuable supplement to the already existing DSM classification. Clustering metrics generated on each distinct stratum (‘Clinical-only’ OR

Fig. 1. Unsupervised learning procedure.
2.4.3. Clustering performance evaluation

Clustering performance and membership were examined using Silhouette scores, and for an exhaustive evaluation, we defined eight other metrics M1-M8 (see supplemental section) which consider various definitions of separability. Regardless of clustering technique, Silhouette scores are widely used for this purpose because they represent how far a point is from its neighboring clusters (Rousseeuw, 1987). If we compute the silhouette scores for a different number of clusters, heuristically we choose the setting that maximizes silhouette score. (Rousseeuw, 1987) By this stated approach, we confirmed that the optimal number of clusters across all our experiments is 3. The silhouette value is a magnitude of how similar an object is to its cluster (cohesion) compared to other clusters (separation), i.e., conversely how dissimilar is one cluster’s center to the points in neighboring clusters. For each object, a certain value \( s(i) \) is identified and plotted. The silhouette scores the object \( i \) is given as:

\[
s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}
\]

\( a(i) \) = average dissimilarity of \( i \) to all other objects of an example cluster \( G1 \),
\( d(i, C) \) = average dissimilarity of \( i \) to all objects of an example neighboring cluster \( G2 \),
\( b(i) \) = minimum \( d(i, C) \).

Silhouette coefficients range from \(-1\) [if \( a(i) < b(i) \)] to \(0\) [if \( a(i) = b(i) \)] to \(+1\) [if \( a(i) > b(i) \)]; i.e. closer to \(-1\) suggesting objects may belong to the wrong cluster, \(0\) indicating minimal separation and positive values closer to \(1\) indicating maximal separation from neighboring clusters. Clustering performance metrics (M1-M8) as shown in Supplementary Table 1 (Refer to appendix for exact formula descriptions mention Supplementary), with lower scores highlighting better separation on all metrics.

2.5. Statistical analysis and cluster validation

All statistical analysis for cluster, evaluation and validation was done using the program R (Vienna, Austria; 2013, http://www.R-project.org).
version 1.1.383). Demographic and clinical characteristics of cluster groups are expressed as means and standard deviations (continuous scales) or as totals and percentages (nominal scales). Group comparisons were performed via analysis of variance (ANOVA) and chi-square tests. For cluster validation, first, contrasts comparing cluster groups (G1, G2 and G3) were run on the composite MRI lobe regions (frontal, parietal, occipital and temporal) of the left and right hemispheres. Adjusted p value (q values) of a comparison was adjusted for the overall number of contrasts across all composite structure with the Hochberg method. (Benjamini and Hochberg, 1995) Similar analysis was run on smooth pursuit latency, gain and accuracy. The Tukey-HSD correction was calculated for the ANOVA contrasting cluster groups on GAF scores and SFS total score. Furthermore, at the proband level, contrasts comparing HCs with each cluster group was run, adjusting for number of overall contrasts with the Hochberg method. The same regions (Frontal, temporal, occipital and parietal) were assessed bilaterally comparing each cluster group to healthy controls. Similar contrasts comparing healthy controls and cluster groups were done with functioning and smooth pursuit scores. An alpha of 0.05 was used throughout.

3. Results

Our unsupervised learning approach as applied on the “composite” data identified three phenotypically distinct clusters. Clinical and demographic characteristics were contrasted between the clusters identified in Table 1. The clusters differed significantly in age, race distribution and family socioeconomic status (Family Hollingshead SES). They also differed significantly in the distribution of DSM IV diagnosis, on the schizo-bipolar spectrum scale (Keshavan et al., 2011) and cluster G1 was generally older. However, there were no significant differences in gender distribution and duration of illness between the three groups.

Clustering results on each separate modality (clinical scores, laboratory markers and composite dataset) is as shown in Figs. 2–5. Each figure depicts a silhouette plot (dotted line depicting average silhouette score for dataset) on the left and a 2-dimension scatter plot on the right (dots with numbers depicting cluster center). Based on results obtained from Clementz et al. (2016) we acknowledge that EEG, saccade, stop signal, and cognition markers used to
construct biotypes are an improvement on existing DSM classification; we therefore used the clustering result of EEG + Saccade + Stop-Signal + Cognition data ("biotype data") as a reference point. We observed a silhouette score of 0.124 using biotype data alone. Clustering the clinical dimensional measures alone after using the non-linear (RBF) kernel improved cluster separation with a silhouette score of 0.29 (Fig. 3). After transforming biotype markers with the non-linear (RBF) kernel and adding categorical DSM diagnosis to biotype markers, we observed a silhouette score of 0.29 (Fig. 3).

Table 1
Demographics and clinical characteristics of probands with psychosis, by ML generated clusters.

<table>
<thead>
<tr>
<th>Region</th>
<th>G1 (n = 190)</th>
<th>G2 (n = 217)</th>
<th>G3 (n = 203)</th>
<th>( \chi^2 )</th>
<th>( p )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
<td>SD</td>
<td></td>
</tr>
<tr>
<td>Duration of illness (Years)</td>
<td>39.02</td>
<td>12.96</td>
<td>36.27</td>
<td>11.72</td>
<td>35.7</td>
</tr>
<tr>
<td>Family Hollingshead score</td>
<td>45.3</td>
<td>16.83</td>
<td>42.58</td>
<td>16.05</td>
<td>38.44</td>
</tr>
<tr>
<td>Schizo-Bipolar scale</td>
<td>5.83</td>
<td>3.07</td>
<td>4.25</td>
<td>3.02</td>
<td>3.96</td>
</tr>
<tr>
<td>Intracranial vol. (cm³)</td>
<td>1405.51</td>
<td>184.75</td>
<td>1432.96</td>
<td>175.99</td>
<td>1467.82</td>
</tr>
<tr>
<td>N% N% N% × 2</td>
<td>120</td>
<td>56</td>
<td>77</td>
<td>35</td>
<td>67</td>
</tr>
</tbody>
</table>

Note: Significant group differences are highlighted in bold letter type.

Note: Structural MRI data available only on G1 (n = 120), G2 (n = 158) and G3 (n = 122) subjects.

Table 2
Significance testing comparing ML generated clusters G1, G2 and G3 comparing cortical thickness.

<table>
<thead>
<tr>
<th>Region</th>
<th>Contrast( a )</th>
<th>d</th>
<th>p</th>
<th>Contrast( b )</th>
<th>d</th>
<th>p</th>
<th>Contrast( c )</th>
<th>d</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frontal L</td>
<td>G1-G2</td>
<td>-0.10</td>
<td>0.39</td>
<td>BT1-BT2</td>
<td>-0.07</td>
<td>0.58</td>
<td>SZ-SZA</td>
<td>0.08</td>
<td>0.53</td>
</tr>
<tr>
<td>Frontal R</td>
<td>G1-G2</td>
<td>-0.36</td>
<td>0.03</td>
<td>BT1-BT2</td>
<td>-0.26</td>
<td>0.09</td>
<td>SZA-BPD</td>
<td>0.31</td>
<td>0.06</td>
</tr>
<tr>
<td>Temporal L</td>
<td>G1-G2</td>
<td>-0.32</td>
<td>0.03</td>
<td>BT1-BT2</td>
<td>-0.31</td>
<td>0.06</td>
<td>SZA-BPD</td>
<td>0.31</td>
<td>0.06</td>
</tr>
<tr>
<td>Temporal R</td>
<td>G1-G2</td>
<td>-0.53</td>
<td>-0.001</td>
<td>BT1-BT2</td>
<td>-0.35</td>
<td>0.02</td>
<td>SZA-BPD</td>
<td>0.27</td>
<td>0.22</td>
</tr>
<tr>
<td>Parietal L</td>
<td>G1-G2</td>
<td>-0.14</td>
<td>0.30</td>
<td>BT1-BT2</td>
<td>-0.02</td>
<td>0.06</td>
<td>SZA-BPD</td>
<td>0.29</td>
<td>0.22</td>
</tr>
<tr>
<td>Parietal R</td>
<td>G1-G2</td>
<td>-0.35</td>
<td>0.02</td>
<td>BT1-BT2</td>
<td>-0.26</td>
<td>0.06</td>
<td>SZ-SZA</td>
<td>0.25</td>
<td>0.22</td>
</tr>
<tr>
<td>Occipital L</td>
<td>G1-G2</td>
<td>-0.32</td>
<td>-0.005</td>
<td>BT1-BT2</td>
<td>-0.31</td>
<td>0.02</td>
<td>SZA-BPD</td>
<td>0.28</td>
<td>0.22</td>
</tr>
<tr>
<td>Occipital R</td>
<td>G1-G2</td>
<td>-0.30</td>
<td>0.06</td>
<td>BT1-BT2</td>
<td>-0.33</td>
<td>0.02</td>
<td>SZA-BPD</td>
<td>0.03</td>
<td>0.83</td>
</tr>
</tbody>
</table>

Note: Significant group differences are highlighted in bold letter type for Cohen’s d effect sizes and Hochberg adjusted p-values.

a Contrasts G1, G2 and G3 derived from clusters generated by composite data (clinical, EEG and cognition).
b Contrasts BT1, BT2 and BT3 derived from clusters generated using EEG and cognition data only i.e. ‘biotypes’ (Clementz et al., 2015).
c Contrasts for DSM-IV classification: SZ (Schizophrenia), SZA (Schizaffective) and BPD (Bipolar disorder).

c.1. Cluster validation with cortical thickness, smooth pursuit, and social functioning

Structural neuroanatomical, smooth pursuit and social functioning data were not used in the creation of clusters and were therefore used as independent means for validating the clustering solutions. Table 2 shows results contrasting G1, G2 and G3 on cortical thickness. G1 is significantly reduced in cortical thickness in the frontal right lobe region compared to G2 and G3, and is significantly lower than G2 in the frontal right lobe. Significantly greater reductions were also observed hemispherically, in the temporal and occipital lobes for G1 and G2 groups, when compared to G3. Probands from the G1 group were most impaired on all pursuit measures of interest compared to controls (Table 2.2). Furthermore, all 3 groups significantly differed in GAF.
scores, with G2 observed to be the worst functioning as assessed by GAF, and G3 doing the best (Table 2.1). Functioning as assessed by SFS scores showed markedly greater impairments in G1 and G2 as compared to G3, with no significant differences between G1 and G2. Table 2 also shows the differences in cortical atrophy between the “biotypes” generated using Clementz et al. and also the DSM IV diagnosis (SZ, SZA and BP).

Fig. 6 shows the deviations of smooth pursuit performance, from that of the healthy subjects and social functioning by cluster groups; Fig. 7 presents effect sizes by region for cluster groups compared to healthy controls. Contrasting with controls, probands within G1 had widespread cortical thickness reductions with a range of much larger magnitudes in effect sizes (d = 0.33–0.64), compared to G2 (d = 0.27–0.52) and G3 (d = 0.02–0.29).

4. Discussion

By using unsupervised learning, we observed cluster separation to be maximal on composite data instead of clinical or laboratory data independently. Inspecting silhouette scores for each phenotype stratum, revealed that three were the optimal number of clusters in each scenario. (Rousseeuw, 1987) With the use of the RBF kernel in transforming the input features, we were better able to separate patients into distinctive subgroups. Refinement in through non-linear methods is illustrated by higher silhouette score of 0.26 (Fig. 4) for clusters formed by clinical dimensional data: compared to clinical data without transformation (Fig. 2). Similarly, laboratory (EEG + Cognition) markers upon transformation achieved a higher silhouette score (0.29). An important feature of our results was that individual silhouette scores of biological and clinical dimensional data are far lesser than that of the composite data clustering (Fig. 5). 6 out of 8 (M1, M3, and M5 through M8) clustering metrics (Supplementary Table 1: Side by side comparison of clustering metrics) had the best discrimination for composite clustering, thereby augmenting the scheme that incorporating a wide variety of data allows selection of optimal clusters for best discerning illness subtypes.

Each cluster contained all DSM diagnoses, with G1 containing a larger proportion of schizophrenia probands (Although 22% were BP and 22% were SZA); G3 had a larger proportion of BP (Although SZ were 33% and SZA were 19%) and group G2 had a fairly even distribution of the 3 disorders (Table 1). Distribution of DSM labels in all three clusters illustrates that most of these disorders have overlaps in clinical manifestations (Keshavan et al., 2013). Even if at times the clinical presentations of these illnesses may vary, the underlying pathophysiology may be similar. Etiological heterogeneity, pleiotropy, and variable expressivity are among other factors, influencing phenotypic manifestations. (R. N. Rosenberg and Pascual, 2014) illustrating the lacunae in relying on clinical symptomatology alone to determine pathognoms of complex psychiatric illnesses.

Another distinguishing feature, is our psychotic subtypes display a cascade in subtype distinctiveness; with subtype G1 showing greatest cortical atrophy in effect size units (Fig. 7), G2 following with slightly milder impairments and G3 showing little to no significant differences when contrasted with healthy controls. G1 likely imitates the prototypical chronic sub group on the psychosis dimension (Keshavan et al., 2013; Tamminga et al., 2013) showing substantive deficits in the cortical thickness. This pattern of deviations was further pronounced in between-subtype contrasts (Table 2: significance testing) where G1 and G2 significantly showed greater cortical thickness deficits from sub-type G3, who were relatively less distinguishable from controls. Our approach demonstrates that there is a notable spectrum of cortical atrophy that spreads across disorders, regardless of their DSM labels. A salient observation to note is the lack of significant difference (p = 0.42) in illness duration between the three subtypes (Table 1), connoting that cortical thickness atrophy maybe related to clinical severity as opposed to disease duration in psychosis (Xiao et al., 2015). Cluster distinctiveness are further highlighted as social functioning scores mirror our cortical thickness findings with G1 and G2 were found highly impaired in social functioning (Table 2.1).

Evidence already suggests that there are large overlaps in symptom based classification methods (Clementz et al., 2016; Pini et al., 2004). By comparing the size and range of effect size differences in Table 2, between our clusters with “Biotypes” and DSM labels, we believe that our approach is a promising in reducing the heterogeneity within psychosis. To the best of our knowledge, very few studies have investigated patterns of abnormalities that differentiate different psychiatric disorders (Bansal et al., 2012; Costafreda et al., 2011; Schnack et al., 2014) albeit in limited one-dimensional approaches (e.g., Imaging, clinical, cognition only). The fact that a stratified dataset combining multiple sources of a patient’s data is much more separable than any individual dataset suggests that there is value in our stratified approach. There are several limitations to our study. Firstly, medication effects could confound classification, as we recruited chronically ill, clinically stable, subjects, most of who were on antipsychotics. Secondly, we only had a single time-point data for every subject. Thirdly, due to missing data, the intersection sample of clinical data, laboratory markers and sMRI data dropped several probands from analysis. However, to our knowledge this is still one of the largest samples of probands applying ML techniques. Fourthly, a common limitation perceived in studies incorporating T1-weighted images for GM thickness, is that the underlying microstructural abnormalities that lead to tissue changes measured cannot be uniquely determined and finally we could not confirm our findings in an independent sample.

Our findings should be taken not as a definitive validation of clinical and biomarker information to define psychotic disorder subtypes, but as a ‘proof of concept’ to investigate the heterogeneity of psychosis. The goal of neuroscience based approaches should focus on identifying biologically homogenous subtypes that cut across phenotypic diagnosis—by sidestepping the issue of a ‘gold standard’ and focusing on ‘stratified psychiatry’ (Kapur et al., 2012). Although, from a translational point of view, many clinics simply do not have the facilities to collect and/or process a whole gamut of clinical and laboratory data which are manually intensive and time-consuming. However, once classification systems in research centers further hone in on efficiency and reliability of
classification techniques, efforts can be undertaken to determine whether reasonably good classification can be performed with fewer and more readily available measures.

Notwithstanding these limitations we believe that our observations at least point towards a promise of combined clinical and biomarker based diagnostics. Further studies need to replicate our observations in larger, more systematically collected samples and also provide further validation of these stratification schemes by collecting longitudinal data as well as etiological data such as molecular gene data.


