Towards Constructing a New Taxonomy for Psychiatry Using Self-reported Symptoms

Jessica Ross a, Thomas Neylan b, Michael Weiner h,c, Linda Chao c, Kristin Samuelson d,e, Ida Sim f

a Department of Psychiatry, University of California, San Francisco, San Francisco, CA, USA
b Department of Psychiatry, University of California, San Francisco and San Francisco VA Medical Center, San Francisco, CA, USA
c Department of Radiology, University of California, San Francisco and San Francisco VA Medical Center, San Francisco, USA
d San Francisco VA Medical Center, San Francisco, CA, USA
e Alliant International University, San Francisco, CA, USA
f Division of General Internal Medicine, Department of Medicine, University of California San Francisco, San Francisco, CA, USA

Abstract

The Diagnostic and Statistical Manual (DSM) has served as the gold standard for psychiatric diagnosis for the past several decades in the USA, and DSM diagnoses mirror mental health and substance abuse diagnoses in ICD-9 and ICD-10. However, DSM diagnoses have severe limitations when used as phenotypes for studies of the pathophysiology underlying mental disorders, as well as for clinical treatment and research. In this paper, we use a novel approach of deconstructing DSM diagnostic criteria, and using expert knowledge to inform feature selection for unsupervised machine learning. We are able to identify clusters of symptoms that stratify subjects with the same DSM disorders into cohorts with increased clinical and biological homogeneity. These findings suggest that itemized self-report symptom data should inform a new taxonomy for psychiatry, and will enhance the bi-directional translation of knowledge from the bench to the clinic through a common terminology.

Keywords:
Psychiatry; Taxonomy; Unsupervised Learning; Clinical and Translational Informatics

Introduction

The Diagnostic and Statistical Manual of Mental Illness (DSM-V) is the current clinical psychiatric classification system in the United States for mental illness and substance abuse disorders [1]. When a patient is given a DSM diagnosis, the clinician asks the patient about the presence or absence of numerous symptoms, and then uses multiple DSM algorithms to determine the best diagnosis for the patient. There is a growing acceptance that DSM syndromes have severe limitations when used as phenotypes for identifying biomarkers associated with mental illness. Whether the symptom data that is ascertained during the diagnostic process can be used more effectively as phenotypes for biological inquiry into the pathophysiology of mental illness remains unclear. The purpose of this study is to determine if atomic level symptom data that is regularly ascertained by mental health practitioners can be effectively used to develop phenotypes that can identify clinical cohorts of psychiatric patients with pathophysiological homogeneity. As these data are already ascertained as part of the normal clinical workflow in psychiatry, and are frequently reported in the medical record, they can theoretically be used to phenotype clinical populations, facilitating cohort identification for biomarker identification, and enabling observational studies on outcomes effectiveness in different cohorts. These studies might in turn lead to the identification of biomarkers that can be used to stratify psychiatric populations in the future, improving diagnosis, treatment, and outcomes in these populations.

The current psychiatric classification system, the DSM, was initially developed after World War II, motivated to a large degree by the need for the Armed Forces to monitor the prevalence of and treat mental illness in soldiers who returned from the War. DSM I in 1952, and DSM II in 1968, were both rooted in psychodynamic psychiatry, in line with the legacy of psychoanalysis (e.g. beginning with Freudian Theory). With DSM III, the decision was made to take an “ atheoretical” approach, and develop a classification system based on patterns of symptoms that clustered together, as agreed to by expert consensus in committee meetings, without the use of empirical data, and divorced from explanatory models of the etiologies of these disorders [2]. Studies showed that this new approach vastly improved inter-rater reliability with respect to the psychiatric diagnosis patients were given [3]. While the approach presumed that the categories reflected underlying pathological processes, no quantitative data or biological markers have been included in the DSM thus far.

Since 1994, the DSM-IV has served as the gold standard in psychiatric treatment and research [4]. Though DSM-5 was officially released in May of 2013, the vast majority of USA clinicians are still transitioning from DSM-IV to DSM-5. As such, DSM-IV disorders have served as phenotypes by which clinicians diagnose and treat patients and obtain reimbursement, as well as phenotypes for basic science, translational, and clinical research carried out in the domain of psychiatric illness. Furthermore, DSM-IV syndromes have been shown to have a great deal of similarity to ICD-9 and ICD-10 diagnoses, so the results of studies using the DSM are highly generalizable to populations worldwide [5].

Much has been written about the limitations of the DSM-IV with respect to the use of its syndromes as phenotypes for use in both clinical practice and research. Below, we will briefly review the most salient points as they relate to phenotypic heterogeneity, limitations of syndromes in identifying biomarkers, and informatics challenges faced when attempting to construct phenotypes that define more clinically and biologically homogenous groups from symptom data ascertained using DSM-IV syndromes as phenotypes.

DSM-IV defines 137 Syndromes/Disorders across 15 categories. Syndromes are by definition binary, and there are no signs (objective or quantitative markers) associated with DSM-IV syndromes. Symptom overlap (e.g. sleep disturbances, guilt) across syndromes (e.g. Major Depressive Disorder (MDD), Post Traumatic Stress Disorder (PTSD)), resulting in clinical presentations where patients are defined as having several “comorbid” psychiatric disorders, complicating
diagnosis and treatment for the clinician, with a paucity of evidence based treatment algorithms for multiple comorbid psychiatric disorders.

DSM-IV phenotypes also have limited clinical utility, as two patients with the same DSM-IV Syndrome may not share any of the same symptoms. Due to the structure of DSM-IV Syndromes, many different combinations of symptoms meet the criteria for the same Syndrome. For example, with Major Depressive Disorder (MDD), “depressed mood” or “anhedonia” AND 5 out of 9 other symptoms are required, leading to 112 different possible symptom presentations. As some of the symptoms are underspecified (e.g. sleep disturbance may refer to hypersomnia or insomnia), it is possible for two patients to have MDD without sharing ANY of the same symptoms. Of note, this lack of syndromic specificity remains present in DSM-5, as many major syndromes, including MDD have not been altered at all for the new version.

This phenotypic heterogeneity also has severe limitations for clinical and translational research. Clinical trials have shown that several classes of medications that are equally efficacious across several syndromes, and are used clinically across several DSMIV higher level classes (e.g. Mood Disorder and Anxiety Disorders, Mood Disorders and Psychotic Disorders) arguing for the fact that the overlap of symptoms across these disorders actually reflects the presence of common pathophysiology present across different DSM-IV syndromes.

There has been an abundance of articles discussing how the lack of biological validity underlying the creation of DSM-IV Syndromes has resulted in a lack of robust biomarker findings within syndromes [6]. In addition to the fact that DSM-IV Syndromes have overlapping symptoms, recent studies have confirmed that specific genetic variants are present in subjects with multiple DSM-IV Syndromes, and not in unaffected controls. One recent seminal study showed that variants in the CACNA1C calcium channel are associated with several psychiatric syndromes, including autism spectrum disorder, attention deficit hyperactivity disorder, bipolar affective disorder, and major depressive disorder [7]. Additionally, several other biological markers, for example non-suppression of cortisol by dexamethasone, have been found to be significantly different in affected individuals as opposed to controls, in multiple psychiatric syndromes [8]. However, to our knowledge, there has not been a comparison of psychiatric symptom presentation similarities across these syndromes, with a focus on how this relates to these biomarker differences.

Finally, the current terminology of the DSM-IV nosology makes it difficult to investigate symptoms across syndromes, due to a decision made with DSM-III that no symptom could be replicated in two Syndromes [9]. Though it is unclear if this rule has explicitly followed during the creation of DSM-IV and 5, manifestations of this limitation undoubtedly exist in DSM-IV and DSM-5, making it challenging at this time to use existing NLP and text-mining methods to automatically identify symptoms across syndromes within narrative text, or even structured interview data. The APA has increasingly published measures to investigate more primary psychiatric symptoms in clinical populations, but the use of these measures has not yet translated into clinical care, nor is it yet reflected in the literature [1].

In view of a preponderance of shortcomings, major stakeholders in mental health treatment and research have publically presented alternate systems for codifying current and future knowledge to inform the understanding of mental illness, most notably Precision Medicine and the Research Domain Criteria (RDoC) [10]. However, these ideas are still very much in the development stage, and there has been a paucity of discussion as to how these taxonomies will be integrated with clinical psychiatry. Generations of psychiatrists, including those who are currently teaching residents and medical students, have been trained using the DSM, and routinely evaluate patients using symptom level data with the ultimate goal of identifying the most appropriate DSM disorder or disorders for the patient based on the constellation of symptoms present in the clinical presentation.

It is with this background that we propose to identify groups of patients that cluster together based on itemized psychiatric symptom level presentation, regardless of how these symptoms have been used to classify DSM Syndromes up until this point. We then propose to identify biomarkers that are associated with these different clinical symptom presentations to show that self-report symptoms have utility in developing a clinically relevant and biologically valid taxonomy for psychiatry. Few studies have looked at the correlation of psychiatric symptoms with biomarkers, independent of DSM-IV syndromes, across a patient population. The vast majority of studies which have looked at itemized symptom level data in the DSM-IV in any manner, have looked only at a subset of symptoms involved in one to three closely related DSM disorders, most notably within childhood disorders such as autism and ADHD, and cognitive disorders including Alzheimer’s. There are two studies we know of that have looked at the DSMIV symptoms in total, but with different foci than described in this proposal [11, 12]. We have not been able to identify any studies that have looked in total at all symptoms across all principal DSM Syndromes within a dataset where patients carry the diagnosis of more than one DSM Syndrome.

Studies have shown that it is possible to classify psychiatric patients into groups of subjects with increased pathophysiological and disease course homogeneity using machine learning algorithms with various types of biomarker data [13]. This classification is possible even without full understanding of the pathophysiology underlying the disorder. These methods also suggest a direct translation from academic research to clinical practice, where they may ultimately be used to diagnose patients in the clinical setting.

We present the analyses of a multi-modal dataset comprised of combat veterans, 33% of whom have been given a DSMIV diagnosis of PTSD. We used unsupervised learning methods, including hierarchical cluster analysis and K means clustering, with the goal of identifying robust groups of patients that cluster together with respect to clinical symptoms and biomarkers.

Materials and Methods

Data

We conducted secondary analyses ascertained through VA and DoD funding at the San Francisco Veterans Administration Health Center. Details of the Gulf War (GW) study have also been published [14, 15]. In brief, there were 292 Gulf War Veterans, 85% of whom were male, and 75% who were combat trauma exposed. Given the small sample size and the relatively large number of clinical and biological markers to be used in the analyses, only the males were included in this study. 50% of those included had comorbid DSM-IV diagnoses of Lifetime Alcohol Abuse or Dependence, and 40% had comorbid DSM-IV diagnoses of Major Depressive Disorder.
The four main data types used in this study are self-report clinical symptom data, and three different modalities of biomarker data: neuropsychological data, magnetic resonance imaging data, and neuroendocrine data. Ascertainment of these data are briefly described below.

The self report clinical data that was used consists of 136 clinical self-reported clinical features, 21 from the Beck Depression Inventory (BDI), a scale developed and used to diagnose Major Depressive Disorder, 19 from the Clinician’s Assessment for PTSD (CAPS), a scale used to diagnose PTSD in clinical populations and 90 items from the Symptom Check List -90 (SCL-90), a clinical questionnaire developed to measure psychiatric symptoms across 10 psychiatric domains, including mood, psychosis, anxiety, eating, and sleep disorders. Also included were 5 binary measures of significant substance use or dependence, for alcohol, cannabinoids, stimulants (e.g. methamphetamine), cocaine, and hallucinogens.

The ascertainment of neuropsychological and imaging data was described in detail in prior publications. Briefly, the neuropsychological battery used assessed three domains of cognitive functioning: verbal memory; visual memory and visual-spatial skills; and attention, working memory, and processing speed with 10 normalized scores reported for each subject. Subjects were studied with MRI; multislice 1H MRSI measures were ascertained. Hippocampal determination was based on MPRAGE images and carried out semiautomatically using a high dimensional brain-mapping tool and values were divided by total intracranial volume for normalization. Subjects also completed a low-dose (0.5 mg) dexamethasone (DST) suppression test challenge using salivary cortisol [8], and area under the curve (AUC) for both day 1 and day 2 of the test were used in these analyses. Baseline serum cortisol level was also ascertained.

**Feature Selection**

A hybrid approach was attempted to reduce redundancy across clinical features: Features were first mapped to each other if a clinical expert (JR) determined the concepts they were identifying as being clinically equivalent (e.g. expected to result in the same answer if both questions were asked to the same patient in a clinical interview). A correlations matrix was also created for all clinical features, to identify features with a correlation coefficient of ≥ 90. Interestingly, there was no intersection between the two methods, and the decision was made to remove a feature based on correlation coefficient ≥ 90, resulting in 124 total clinical features that were ultimately used for the subsequent analyses.

Ideally we would have been able to use symptom data from the Structured Interview for DSM (SCID), the tool that was used to initially identify DSM Diagnoses. However, the data in this study has been generated without the SCID data, as we only currently have composite level syndrome data available from the SCID. We therefore mapped the complete CAPS, BDI, SCL-90, and Alcoholic and substance abuse portions of the SCID to all the SCID questions, and estimate that our data cover approximately 85 % of all symptoms ascertained in a formal SCID, with predominantly 75% of psychotic symptom data and 50% of OCD symptom data unaccounted for. However, as individuals were excluded from our study with a history of psychotic symptoms, we estimate that missing relevant symptomatic data are minimal.

**Statistical Analyses**

Individuals with > 15% data missing in any measures (clinical, imaging, endocrine, neuropsychiatric), were not included for any analyses that required that data type. Missing values were filled with median imputation. All data used in the analyses was standardized and normalized.

All statistical analyses were performed using R statistical programming software [16]. We used hierarchical clustering (hclust) with Ward’s distance measurement. As multiscale bootstrap resampling (pvclust) provided no significant p-values (e.g. p<.05) for any individual clusters, clusters were delineated using a dendrogram to visualize vertical distances in branching across clusters. We then performed K-means nearest (kmeans) neighbor clustering three different times, as clusters delineated with the kmeans algorithm are nondeterministic. We identified the optimal number of clusters from the K-means analyses by finding the bend in the plot of the within group sum of squares versus number of clusters. We then compared cluster solutions across all three K-means results with the hierarchical clustering results to determine robustness of clusters. Finally, we performed descriptive statistical analysis on imaging, neuropsychiatric, and neuroendocrine measures using ANCOVA to determine differences in biomarkers across clusters.

**Results**

Hierarchical agglomerative clustering using 124 features using Ward’s method for determining distance between individuals produced the results shown in Figure 1. Visual inspection of dendrogram branching identified two, three, or five natural clusters. Multiscale bootstrap resampling provided no significant p-values (e.g. p<.05) for any individual clusters. Three independent runs of K-nearest neighbor analysis also revealed that data ideally clustered into 5 groups that had a good to great degree of similarity with the hierarchical clusters (Cohen’s Kappa Values .904, .897, .702).

Given the similarity of the multiple solutions and the determinism in using agglomerative hierarchical clustering, the five clusters delineated through hierarchical clustering were chosen for further analyses of biomarkers across clusters. Ancova with adjustment for age was calculated for all biomarkers in the data set. The most significant p-values, values without adjustment for multiple testing are shown in Table 1. In an effort to account for variance, a linear model was constructed using all variables that had unadjusted p-values approaching a level of significance in a Mannova analysis, with results for differences in biomarkers across clusters shown in Table 2.

To further gain insight into differences between the five clusters, we looked at the absolute values of individual symptoms across the five clusters, as shown in Figures 2a, 2b, and 2c. We also graphed mean values of all biomarkers across the five clusters to facilitate description of these five clusters both from the perspective of biomarkers, and of clinical symptoms. Graphs for the six biomarker variables that showed significance in the Manova analyses, are shown in Figures 3a through 3f.
Discussion

In this study we show that statistical learning methods can be used with atomic level psychiatric symptom data to identify groups of patients with phenotypic and biological homogeneity. Perhaps the most surprising finding is that the clusters delineated are not associated with any “type” or “types” of symptoms as would have been expected from the DSM classification system. Instead, the 5 clusters are defined by the intensity of the majority of symptoms across all questionnaires, and hence, DSM Syndromes. In fact, with the exception of two neuro-vegetative symptoms associated with depression and one paranoia symptom, the subjects in Cluster 1 on average have worse values for all of the psychiatric symptoms than subjects in all of the other clusters. As would be expected, these subjects have a higher mean number of DSM–IV diagnoses, yet no single or even group of co-morbid diagnoses can partition them in the intuitive manner that hierarchical clustering is able to.

Cluster analyses delineated 3 clusters of patients that each contain significant numbers of individuals who meet criteria for PTSD (clusters 1, 2, and 4), as seen in Table 3. However, individuals in cluster 4 have much lower levels of subjective psychiatric symptoms than individuals in clusters 1 or 2, and significantly different (and arguably more “normal”) biological markers across the neuropsychiatric data, imaging data, and neuroendocrine data. These findings are in line with the growing consensus that DSM phenotypes are underspecified, and lump together groups of patients that are too heterogeneous to use in clinical, translational, or basic science studies. While this study is obviously limited by the very small sample size and exclusion of female subjects, the fact that biologically and clinically relevant clusters were identified, provides support for the hypothesis that using data from larger psychiatric cohorts with these methods will greatly increase our current knowledge base in the psychiatric domain.

The use of DSM diagnoses does not correlate with these clusters, but symptom data analogous to that used to delineate DSM diagnoses were used in these analyses. Therefore, these findings propose a method by which clinical stakeholders and basic science researchers in the mental health domain can communicate to delineate a taxonomy for psychiatry that is ultimately both clinically and biologically useful. Arguably, phenotype definition is the most pressing issue in psychiatry.

### Table 1- Ancova of Clusters with Significant Biomarkers

<table>
<thead>
<tr>
<th>Dependent Variable</th>
<th>Ancova p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Right Hippocampal Volume</td>
<td>.057</td>
</tr>
<tr>
<td>Dexamethasone Suppression Test</td>
<td>.004</td>
</tr>
<tr>
<td>Executive Functioning</td>
<td>.057</td>
</tr>
<tr>
<td>Processing Speed</td>
<td>.030</td>
</tr>
<tr>
<td>Manual Dexterity</td>
<td>.007</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>.016</td>
</tr>
</tbody>
</table>

### Table 2- Mancova of Clusters using Significant Biomarkers

<table>
<thead>
<tr>
<th>Biomarker</th>
<th>Mancova p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>DST:Area Under the Curve day 1</td>
<td>.910</td>
</tr>
<tr>
<td>DST:Area Under the Curve day 2</td>
<td>.008</td>
</tr>
<tr>
<td>Baseline Serum Cortisol</td>
<td>.222</td>
</tr>
<tr>
<td>Memory</td>
<td>.024</td>
</tr>
<tr>
<td>Executive Functioning</td>
<td>.003</td>
</tr>
<tr>
<td>Learning</td>
<td>.039</td>
</tr>
<tr>
<td>Attention</td>
<td>.034</td>
</tr>
<tr>
<td>Processing Speed</td>
<td>.003</td>
</tr>
<tr>
<td>Manual Dexterity</td>
<td>.056</td>
</tr>
<tr>
<td>Ataxia</td>
<td>.004</td>
</tr>
<tr>
<td>Verbal IQ</td>
<td>.003</td>
</tr>
<tr>
<td>Performance IQ</td>
<td>.000</td>
</tr>
<tr>
<td>Right Hippocampal Volume</td>
<td>.417</td>
</tr>
<tr>
<td>Left Hippocampal Volume</td>
<td>.517</td>
</tr>
</tbody>
</table>

Figure 2- Mean itemized symptom data for each cluster

Figure 3- Normalized Mean Biomarker Values By Cluster: a) Right Hippocampal Volume, b) DST Area Under the Curve, c-e) Standardized Executive Functioning, Processing Speed, and Manual Dexterity, f) Performance IQ
to date. DSM-5, released in May 2013, remains the clinical bible for psychiatrists, and its diagnoses, which map to ICD-9 and ICD-10, are necessary for billing insurance companies and Medicare. DSM-5 therefore remains the standard in medicine, and continues to be taught to medical students and residents. Yet one month prior to the release of DSM-5, the NIMH proclaimed that studies using DSM phenotypes will no longer be a priority for funding, strongly encouraging the use of the RDoCs [10]. While a biologically based taxonomy is the ultimate goal for psychiatry, as with all other domains of medicine, clinical psychiatry is far from being grounded by mechanistic biomarkers. The findings in this paper that individual symptoms that are regularly ascertained clinically are associated with changes across multi-modal biomarkers, gives support to the notion that symptoms may still be useful in psychiatry to both clinicians and researchers. In fact, there may be data within the multitude of large datasets already ascertained, that can be used to develop symptom based psychiatric phenotypes that can identify clinically and biologically homogeneous subpopulations, both for diagnostic purposes, and that will have clinical utility and be able to further inform phenotype construction in a biologically based taxonomy.

It seems likely in the near future that patients will continue to present with self-report symptoms that will guide the evaluation and treatment of patients. While currently clinicians work towards identifying the most appropriate DSM diagnosis or diagnoses for each client to guide treatment, it may require few extra resources to obtain more extensive detailed symptom level data, that can be used with statistical learning methods to cluster clients into groups with increased biological and clinical homogeneity. These cohorts can be followed in observational studies in outcomes effectiveness, and potentially be recruited to participate in biomarker studies. Additionally, in the age of the Internet, ascertaining self-report symptoms may require few extra resources to obtain more extensive detailed symptom level data, that can be used with statistical learning methods to cluster clients into groups with increased biological and clinical homogeneity. The findings in this paper that individual symptoms that are regularly ascertained clinically are associated with changes across multi-modal biomarkers, gives support to the notion that symptoms may still be useful in psychiatry to both clinicians and researchers. In fact, there may be data within the multitude of large datasets already ascertained, that can be used to develop symptom based psychiatric phenotypes that can identify clinically and biologically homogeneous subpopulations, both for diagnostic purposes, and that will have clinical utility and be able to further inform phenotype construction in a biologically based taxonomy.

Acknowledgments

Jessica Ross was supported by the Veteran’s Administration Advanced Fellowship in Psychiatry/Neurosciences.

References


Address for correspondence

J.A.Ross. Assistant Clinical Professor, Department of Psychiatry, University of California, San Francisco 1001 Potrero Avenue, 7M, San Francisco, CA, 94110. Email: Jessica.Ross@ucsf.edu