Characterizing Data-Driven Phenotypes of Schizophrenia, Bipolar, and Attention Deficit/Hyperactivity Disorder

Scott L. Fleming

1 Introduction

The field of clinical psychiatry is undergoing a radical transformation in the way it characterizes and diagnoses mental health disorders. Whereas traditionally psychiatrists have relied on the Diagnostic Statistical Manual of Mental Disorders 5 (DSM-5) to make diagnoses, the manual has been criticized in recent years because boundaries between disorders are not as strict as the DSM suggests [Casey et al., 2013]. Indeed, it is likely that the categories in the DSM conflate several types of individuals in whom diverse brain dysfunctions drive symptoms [Williams, 2016]. The challenge, then, is to identify biomarkers that correlate with underlying brain dysfunction [Drysdale et al., 2017].

One promising hypothesis suggests that, once these biomarkers are identified, patients could be stratified into neurobiologically meaningful clusters, each of which would represent an underlying type of brain dysfunction [Williams, 2016]. Identifying these biomarkers and their associated brain dysfunction categories could lead to more neurobiology-based diagnoses, which would spur the development of targeted interventions and increase the effectiveness of mental health treatment overall. The aims of this project are therefore to 1) identify information-rich biomarkers in a cohort of healthy participants and patients with a variety mental health disorders; 2) assess the degree to which stratification based on these biomarkers correspond to current DSM disease categories; 3) evaluate the hypothesis that individuals with identified mental health disorders tend to form clusters based on these biomarkers.

This analysis leveraged data from the Consortium for Neuropsychiatric Phenomics, which collected performance metrics on a suite neurocognitive tasks from 50 subjects with schizophrenia, 49 subjects with bipolar disorder, 43 subjects with attention deficit/hyperactivity disorder (ADHD), and 130 healthy individuals to achieve the aims stated above. The inputs in this case were the subjects' performance on these neurocognitive tasks, and the outputs are both a set of latent features that describe these subjects in a reduced feature space and a clustering based on those features.

2 Related Work

Interest in data-driven methods in Psychiatry have increased substantially over the last several years [Biswal et al., 2010, Akil et al., 2011]. Many research efforts have focused on discovering heterogeneity and subtypes within mental health disorders [Wardenaar and de Jonge, 2013]. Numerous studies in depression, for example, have used factor analytic methods, principal components analysis, and latent class analysis on self-report symptom and imaging data to discover subtypes of depression disorders, with varying degrees of success [Van Loo et al., 2012]. There have also been some efforts to characterize similarities and differences between diagnoses using brain imaging; Sui et al. [2011] found that patients with schizophrenia differed significantly from patients with bipolar disorder based on functional activity and white matter integrity in certain brain regions. Arribas et al. [2010] used a generalized softmax perceptron neural network to discriminate between healthy controls, bipolar disorder, and patients with schizophrenia and were able to an accuracy of 70%.

While the accuracy they achieved using just neuroimaging features in the study by [Arribas et al., 2010] was perhaps promising, one critique often levied against such purely supervised analyses is the issue of "reification" [Hyman, 2010]. That is to say, as the field tries to move away from the symptom-based diagnostic labels established in the DSM-V and toward a more neurobiologically based science, studies that base the success of their analysis entirely on predicting DSM-V labels may just be discovering noise that coincidentally predicts poorly-specified targets. Or, under the assumption that the DSM-V diagnostic labels are not entirely useless, some might argue that data-driven analyses should focus first on establishing concretely the "right" diagnostic labels and only then begin training classifiers to distinguish between them. Thus, the methods in this paper focus predominantly on unsupervised rather than supervised learning techniques. Additionally, while neuroimaging is satisfying in the sense that it assays brain activity directly, other aspects of cognition displayed in performance on cognitive tasks may capture different aspects of underlying neural dysfunction and cost much less as a diagnostic tool. A primary contribution of this paper is the first documented use of a variational autoencoder to generate a low-dimensional representation of subjects based on neurocognitive task data alone. We hope this will serve as a template for further analyses attempting to discover and characterize latent features underlying psychiatric disorders.

3 Dataset and Features

A recent study from the Consortium for Neuropsychiatric Phenomics (CNP) made public a promising dataset containing the results from extensive neuropsychological and neuroimaging assessments on over 200 participants [Poldrack et al., 2016]. The CNP dataset was collected as part of an effort to clarify phenotype domains, particularly those of memory/working memory and response inhibition, which are implicated in bipolar disorder, schizophrenia, and ADHD. To that end, researchers collected interviews and rating scales, self-report measures, neurocognitive exams [...], and neuroimaging data. [Poldrack et al., 2016]. The behavioral assessments are divided into domains of symptoms, traits, neurocognitive tasks, and neuropsychological assessments. For this project, we focus on the behavioral assessments and leave the neuroimaging features as a future opportunity for research. Roughly half of the participants had a diagnosis for one or more mental health disorders from the DSM-5 and the other half were healthy individuals (see Table I).

TABLE I: *Breakdown of the CNP Cohort*

Variable Name	Original Dataset	Train Set	Validation Set	Test Set
Control	130	78	26	25
ADHD	43	16	4	5
Bipolar	49	25	8	12
Schizophrenia	50	25	10	7
Total	**272**	**144**	**48**	**49**

While the CNP dataset has myriad features, not all of them are relevant. Many features are overlapping, many have near zero variance, a number have a high degree of missingness and even more are highly correlated. While columns with zero variance can be reasonably excluded without trouble, eliminating features with missingness can be more problematic, as one can lose potentially important features or diminish the sample size of the study if we blindly censor rows and columns with missing data. The original CNP dataset had 272 participants and 1800 behavioral assessments/features.

In choosing the initial set of features to include our analysis, we excluded all features with zero variance (i.e. features in which there was only one value for all of the participants in the study) and we excluded features that had more than 40% missing values each (i.e. we removed features that were $< 60\%$ complete). After reducing the number of features/columns, we kept only those participants with at most 12 missing values. Missing measurements for the remaining participants were imputed using K-Nearest Neighbor Imputation, in which a missing value j for participant i is assigned to be the average value of j measured across the 5 nearest data points/"neighbors" to participant i (here proximity is measured as the euclidean distance between participants over all features).

Finally, based on this cleaned cohort, we split the data into a training set with 144 participants (60% of the cohort), a validation set with 48 participants (20% of the cohort), and a test set with 49 participants (20% of the cohort). The breakdown of the final cohort by DSM-V diagnosis (including the breakdown for the train, validation, and test sets) is shown in Table I. The preprocessing left 1320 features, of which 375 were neurocognitive tasks and neuropsychological assessments. The latter were used as a the primary basis of our analysis.

4 Methods

4.1 Variational Autoencoder

Our goal of characterizing data-driven phenotypes for schizophrenia, bipolar disorder, and ADHD lends itself well to Bayesian analysis and variational inference. In psychiatry, it is understood that the suite of behaviors displayed by an individual are generated from underlying neurocognitive processes. If we wish to make biologically-driven diagnoses, we must focus our efforts and analysis on modeling those neurocognitive processes. Indeed, any unsupervised learning (e.g. clustering) would ideally be performed on those latent features, if we could somehow accurately model them. Given that those neurocognitive processes are not directly observable, however, our goal becomes one of modeling those processes as latent variables and developing posterior estimates for them given the observable data in our dataset. Variational Bayesian approaches allows one to approximate these latent posteriors. In particular, the Auto-Encoding Variational Bayes (AEVB) allows one to approximate the posterior of continuous latent variables, when it is assumed that the dataset is i.i.d. and each datapoint is generated from those continuous latent variables (as we would assume in our dataset) [Kingma and Welling, 2013].

Generally, autoencoders take an input, encode that input into a lower-dimensional space by way of a neural network, then decode the lower-dimensional representation back to the original input. As autoencoders are trained, they can generate more and more faithful representations of the original input via the low-dimensional latent feature space. Variational autoencoders (VAEs) take a twist on this idea in that they "push" the latent variable representation to take on a standard Normal distribution. This is accomplished efficiently by utilizing a specialized loss function and incorporating stochasticity into the encoding process. Specifically, if x is the raw feature representation of a subject in our dataset, then the VAE uses the neural net to encode x into two latent vectors, z_μ and z_σ, such that the encoding is given by

$$h = g(Wx) \tag{1}$$
$$z_\mu = W_\mu h \tag{2}$$
$$z_\sigma = W_\sigma h \tag{3}$$
$$q_\theta(z|x) = z_\mu + z_\sigma \odot \epsilon \tag{4}$$
$$h' = g'(W_\psi q_\theta(z|x)) \tag{5}$$
$$p_\psi(\tilde{x}|z) = sigmoid(h') \tag{6}$$

where W represents the transformation of the data onto an intermediate hidden layer, h, (followed by a nonlinear transformation g), W_μ and W_σ represent the linear transformation of h into the latent space, and $\epsilon \sim \mathcal{N}(0, I)$ represents random noise. The latent variable representation thus takes on a gaussian form which is propagated through another

hidden layer of the neural network (represented by the linear transformation W_ψ followed by a nonlinear transformation g') to decode the latent variable into a representation of the original input, $\tilde{x}$. The loss function is then given by

$$L^{(i)}(\theta, \psi) = \frac{1}{m} \sum_{i=1}^{m} \left(-E_{z \sim q_\theta(z|x^{(i)})} \left[\log p_\psi(x^{(i)}) \right] + KL \left(q_\theta(z|x^{(i)}) || p(z) \right) \right) \tag{7}$$

where the first term in the sum represents the reconstruction loss and the second term represents the Kullback-Leibler divergence between the posterior of the latent variable and the marginal distribution of the latent variable, which we would like be standard Normal. As we backpropagate the loss through the network, the second term encourages the latent representation to take on a standard normal form, which avoids issues of arbitrary encoding by the VAE and encourages the autoencoder to maintain a "meaningful" sense of distance between the points in our dataset [jaa]. This method was used to generate latent representations of our dataset, which can be seen in Figure 3.

4.2 Principal Components Analysis

As a way to compare the novel VAE encoding with a technique more commonly used in the field of Psychiatry, we also performed Principal Components Analysis on my data (after centering and scaling the data to have mean 0 and variance 1, i.e. $\sigma^2 = 1$) in order to create a lower-dimensional representation of each subject. This was accomplished by finding the eigenvalues of the empirical covariance matrix, Σ. Specifically, the implementation in `R` that we used finds the low-rank representation of X by diagonalizing the covariance matrix and taking the SVD of the original data matrix, such that

$$X = USV^T \tag{8}$$

where X is the original data matrix (centered and scaled), US represents the scores of the PCA (the datapoints represented by principal components rather than the original features) and V represents the loadings of the original features onto the principal components. The representation of the participants in this lower dimensional space is shown in Figure 3.

4.3 Hierarchical Clustering

A common conceptualization of mental health disorders is that there are broad classes of disorders, each of which can be subdivided into specific types of disorders. In this sense, agglomerative hierarchical clustering seemed like a natural way to model the data. Agglomerative clustering starts with individual points and iteratively combines clusters based on some distance criteria and linkage heuristic. Disease phenotypes that are more similar would therefore be represented as more closely related in the resulting clustering, with subtypes of disease still nested together. We used correlation as a distance metric under the auspice that patients should be grouped by profiles of cognition rather than severity of symptoms. That is to say, if $d_{i,j}$ were the distance between subject $x^{(i)}$ and subject $x^{(j)}$, then we would have

$$d(x^{(i)}, x^{(j)}) = 1 - corr(x^{(i)}, x^{(j)}) \tag{9}$$

Additionally, we used complete linkage, a heuristic which states that the next clusters to merge/agglomerate at any given point are those with the minimum maximal distance between points within their clusters. In other words, one would merge C_i and C_j that minimize

$$D(C_i, C_j) = \max_{x^{(i)} \in C_i, x^{(j)} \in C_j} d(x^{(i)}, x^{(j)}) \tag{10}$$

4.4 Choosing the Number of Clusters via the Gap Statistic

Hierarchical clustering produces a large number of potential clusterings (depending on where one cuts the resulting dendrogram), however it does not necessarily offer any insight into which number of clusters is "optimal". To address this problem, Tibshirani et al. [2001] developed a metric and heuristic for estimating the optimal number of clusters. Specifically, they developed the Gap Statistic, which is essentially an approximation of how different the pooled within-cluster sum of squares you obtain on your actual dataset differs from what you would expect under a null distribution. More specifically, if D_r is the sum of all pairwise distances in cluster C_r, then

$$W_k = \sum_{r=1}^{k} \frac{1}{2|C_r|} D_r \tag{11}$$

represents the pooled within-cluster sum of squares around the cluster means, calculated over all clusters in the dataset, and the Gap Statistic is

$$Gap_n(k) = \frac{1}{B} \sum_b (\log W^*_{kb}) - \log W_k \tag{12}$$

where W^*_{kb} is calculated by generating a uniform distribution over the feature space of the original data, then performing hierarchical clustering on that uniform/null distribution and cutting the dendrogram at the point associated with k clusters. The Gap statistic will be high when the pooled within-cluster sum of squares in your actual data is much less than the pooled within-cluster sum of squares you would expect by chance [Tibshirani et al., 2001]. Thus, a reasonable heuristic is to choose the number of clusters associated with a first local maximum of the Gap statistic, as that represents the point at which partitioning the data further does not lead to pooled within-cluster variance. We used 50 bootstrap samples for $k = 1, 2, \ldots, 7$ clusters and calculated the Gap statistic at each point.

5 Results and Discussion

We chose to use 20 latent variables to represent my dataset using the Virtual Autoencoder. This was approximately in concordance with current theories surrounding the number of factors that contribute to psychiatric phenotypes seen in bipolar, schizophrenia, and ADHD [Van Dam et al., 2017]. The VAE was trained using stochastic gradient descent, with 150 epochs and a minibatch size of 33. Given the small size of the dataset, computational time was not an issue and so we chose a minibatch size that was simply half the training set. Indeed, the loss seemed to taper off with this number of epochs and minibatch size, suggesting a reasonable fit (see 1). Even with that compression, we were able to find a representation that resembled the PCA results for the first two principal components (see Figure 3. This was promising in the sense that it suggested the difference seen between subjects with Adult ADHD and schizophrenia were consistently separated in both representations of the data.

Performing the hierarchical clustering as mentioned previously and calculating the Gap statistic for $k = 1, 2, \ldots 7$ clusters we found that the Gap Statistic was highest when clustering on the latent features from the VAE for 3 clusters (see figure 2). The clustering that resulted can be seen in 4) and the associated table representing cluster membership by disorder is shown in Table II.

Note that, although clustering is decidedly an unsupervised learning method, if we were to classify the participants based on cluster membership we would have an accuracy of 0.51, with sensitivities of ADHD, Bipolar, and Schizophrenia being 0.6875, 0.48, and 0.44 respectively. Here the sensitivities are calculated as

$$\text{Sensitivity for category } k = \frac{\text{Number of subjects predicted to have diagnosis k that actually have diagnosis k}}{\text{Total number of subjects known to actually have diagnosis k}} \tag{13}$$

This is actually in accordance with our current understanding of Bipolar and Schizophrenia, namely that the two are poorly distinguished within the DSM-V and may in fact be overlapping significantly in terms of their symptom profiles, while ADHD tends to be more distinct. This suggests that there may be new categories of diagnosis that group certain bipolar and schizophrenia patients together into one subtype and other bipolar and schizophrenia patients together into a separate subtype. Further analysis is needed to understand exactly which features are over- and under-expressed in cluster 2 relative to cluster 3.

Additionally, note that ADHD was clustered into a group altogether quite separate from Schizophrenia. This is also in line with current understanding of symptom profiles. One hypothesis could be that those bipolar patients grouped with the ADHD patients were hypermanic rather than hypomanic (which is typically more associated with Schizophrenia).

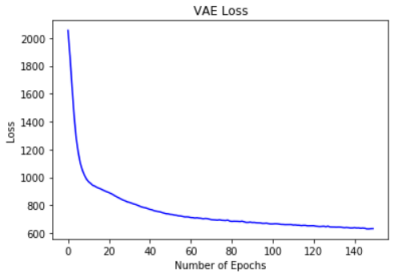

Figure 1: Loss over the number of epochs for the training of the VAE. Note that the loss tapers off by 150 epochs.

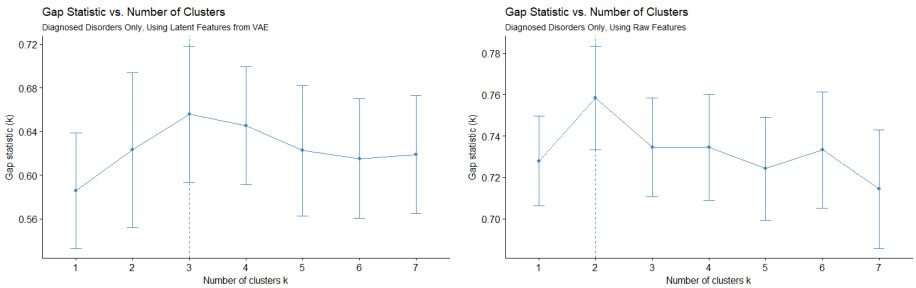

Figure 2: Gap statistic for clustering on both the latent features generated from the VAE and from the raw data. Clustering on the latent features suggests 3 clusters, while clustering on the raw features suggests only 2.

6 Conclusion

In summary, this exploratory analysis offers a method with great potential for discovering subtypes of disease not currently extant in the current DSM-V categorization but that nevertheless capture objective underlying neurological measures. Further work is needed to characterize the underlying latent features. Specifically, cross-correlations between the latent features and symptom/trait features in the dataset could provide insight into which types of symptoms predominate in

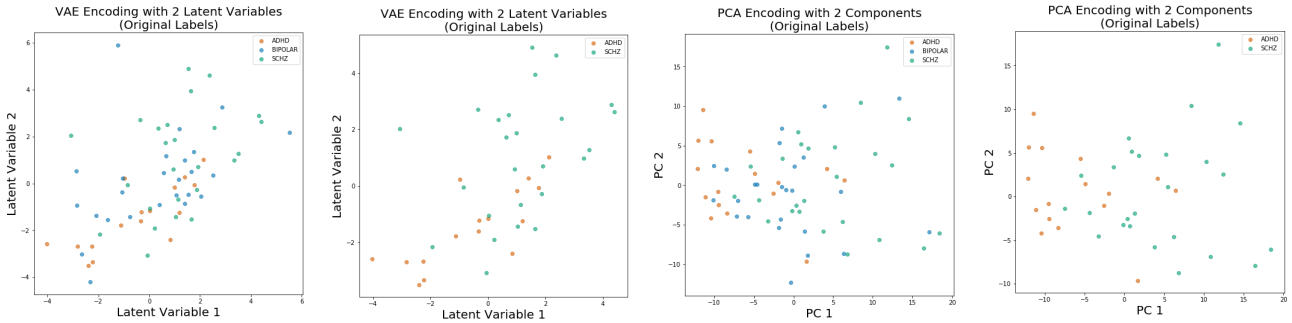

Figure 3: Latent variable representation using both VAE and PCA. Note that while the representation seems unintelligible when bipolar disorder subjects are included, the distinction between ADHD and Schizophrenia is actually quite marked.

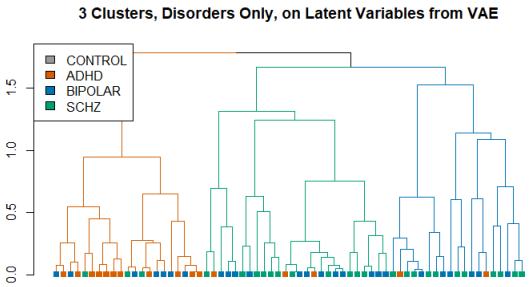

Figure 4: Dendrogram resulting from clustering on the Latent Features from the VAE (left) and the raw features (right).

each of the clusters. Additionally, the validation and test set were set aside for further analysis that will attempt to classify participants based on their diagnosis, comparing those accuracies with the "accuracies" obtained from clustering alone. We are optimistic that this work will result in a deeper understanding of psychiatric disorders and their associated biology.

TABLE II: *Cluster Result for Hierarchical Clustering on VAE Latent Variables*

	ADHD	**Bipolar**	**Schizophrenia**
Cluster 1	11	7	3
Cluster 2	3	12	11
Cluster 3	2	6	11

7 Contributions/Acknowledgements

I did not do this project with a team, but I did have oversight from Dr. Russell Poldrack in the Psychiatry Department. His contribution was essentially that of pointing me to the dataset and making me aware of variational autoencoders. But all data gathering, cleaning, preprocessing, and analysis are my own work (our collaboration, in total, maybe spanned the course of 3 hours. Maybe.)

References

Huda Akil, Maryann E Martone, and David C Van Essen. Challenges and opportunities in mining neuroscience data. *science*, 331(6018):708–712, 2011.

Juan I Arribas, Vince D Calhoun, and Tülay Adali. Automatic bayesian classification of healthy controls, bipolar disorder, and schizophrenia using intrinsic connectivity maps from fmri data. *IEEE Transactions on Biomedical Engineering*, 57 (12):2850–2860, 2010.

Bharat B Biswal, Maarten Mennes, Xi-Nian Zuo, Suril Gohel, Clare Kelly, Steve M Smith, Christian F Beckmann, Jonathan S Adelstein, Randy L Buckner, Stan Colcombe, et al. Toward discovery science of human brain function. *Proceedings of the National Academy of Sciences*, 107(10):4734–4739, 2010.

BJ Casey, Nick Craddock, Bruce N Cuthbert, Steven E Hyman, Francis S Lee, and Kerry J Ressler. Dsm-5 and rdoc: progress in psychiatry research? *Nature Reviews Neuroscience*, 14(11):810–814, 2013.

Andrew T Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N Fetcho, Benjamin Zebley, Desmond J Oathes, Amit Etkin, et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature medicine*, 23(1):28–38, 2017.

Steven E Hyman. The diagnosis of mental disorders: the problem of reification. *Annual review of clinical psychology*, 6: 155–179, 2010.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

RA Poldrack, Eliza Congdon, William Triplett, KJ Gorgolewski, KH Karlsgodt, JA Mumford, FW Sabb, NB Freimer, ED London, TD Cannon, et al. A phenome-wide examination of neural and cognitive function. *Scientific data*, 3: 160110, 2016.

Jing Sui, Godfrey Pearlson, Arvind Caprihan, Tülay Adali, Kent A Kiehl, Jingyu Liu, Jeremy Yamamoto, and Vince D Calhoun. Discriminating schizophrenia and bipolar disorder by fusing fmri and dti in a multimodal cca+ joint ica model. *Neuroimage*, 57(3):839–855, 2011.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Nicholas T Van Dam, David OConnor, Enitan T Marcelle, Erica J Ho, R Cameron Craddock, Russell H Tobe, Vilma Gabbay, James J Hudziak, F Xavier Castellanos, Bennett L Leventhal, et al. Data-driven phenotypic categorization for neurobiological analyses: Beyond dsm-5 labels. *Biological psychiatry*, 81(6):484–494, 2017.

Hanna M Van Loo, Peter De Jonge, Jan-Willem Romeijn, Ronald C Kessler, and Robert A Schoevers. Data-driven subtypes of major depressive disorder: a systematic review. *BMC medicine*, 10(1):156, 2012.

Klaas J Wardenaar and Peter de Jonge. Diagnostic heterogeneity in psychiatry: towards an empirical solution. *BMC medicine*, 11(1):201, 2013.

Leanne M Williams. Precision psychiatry: a neural circuit taxonomy for depression and anxiety. *The Lancet Psychiatry*, 3(5):472–480, 2016.