The Genetic Architecture of Psychiatric Illness

Authors: Ziwei Cheng, Jack Lovell, Frida Muedsam

Date: 05/03/2021

Abstract

Mood and psychiatric disorders are associated with people's everyday mental health, emotional states, and life circumstances. As scientists continue to investigate the onsets, causes, and treatment of different mood and psychiatric outcomes, different and diverse techniques have also evolved and become more powerful. This paper aims to understand the shared genetic risk between three mood disorders: major depressive disorder, substance-use disorder, and post-traumatic stress disorder. To answer this question, we may use a relatively new technique called genomic structural equation modeling (gSEM). For the gSEM model, we curate GWAS summary stats for all the phenotypes we are interested in. A GWAS is a study that identifies genetic variants in a population of interest vs. healthy controls. We found that the common factor model using diagonally weighted least square did not yield any significant results. Three specific models used each of the traits as the outcome and the other two as the predictors. The specific models didn't find any significant association between predictor traits and outcome traits. Additionally, we found that in the standardized cases, the correlation between SUD and PTSD is significant ($p<0.05$), the correlation between MDD and PTSD is significant ($p<0.05$), and the correlation between MDD and SUD is not significant ($p>0.05$).

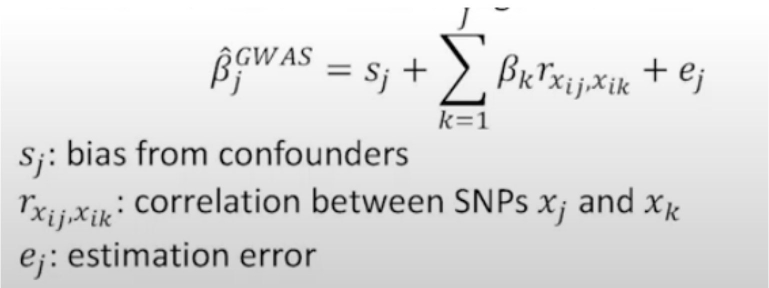
Introduction

Mood disorders and psychiatric conditions are complex illnesses that impact a large proportion of the population. In 2021, the Anxiety and Depression Association of America reported that 16.1 million adults are affected by Major Depressive Disorder (MDD), 7.7 million adults are affected by Post Traumatic Stress Disorder (PTSD), and 24 percent of people prescribed opiates misuse them, resulting in a Substance Use Disorder (SUD). The treatment and even diagnosis of these illnesses are largely still under development for many reasons. The main reason being a striking heterogeneity of the associated disorders' etiology and pathology (Kennedy et al., 2019). Despite this heterogeneity, the sequencing of the entire human genome provided promise for understanding the genetic basis of mood and psychiatric conditions. While their more physiological counterparts in the medical world have significantly benefited from the advancement of our understanding of the human genome, the genetic basis of many behavioral health disorders still is not apparent. The reasons for this are not entirely clear, and the heterogeneity is undoubtedly attributable to this fact.

After many studies and a considerable effort to determine the genetic sequencing of the mood and psychiatric disorders mentioned above, a picture emerged that is much more complex. With a striking amount of heterogeneity observed across GWAS within a disease, in addition to shared variants between certain diseases. This heterogeneity is realized in the results from each different model, which is signified by single nucleotide polymorphisms (SNPs) differing across studies. These SNPs are observed in an individual's DNA, and it's simply a deviation of the base-pair from the healthy or control population in question (Manolio 2010). This problem is quite apparent in its nature. If a clinician wants to develop a gene therapy for MDD, which of the hundreds if not thousands of genes do they choose from? What might this mean for someone suffering from more than one disease, such as MDD and a SUD? A recent technique was developed called "genomic structural equation modeling" (gSEM) that allows scientists to investigate this heterogeneity and shared genetic architecture to solve this problem. To model this relationship between and within mood disorders, an investigator must curate GWAS summary statistics for the phenotypes they are interested in and submit them to their model (Grotzinger, A.D., et al. 2019). In this paper, we plan on curating GWAS summary statistics for MDD, PTSD, and SUDs which will be submitted to a gSEM.

After a GWAS is performed, summary statistics are available in the form of tabulated data, where each SNP in the genome has a particular p-value, indicating whether it is significant in the current model. The typical model used is a multi-regression linear model with the behavioral phenotype of interest as the outcome and a section of the genome as the predictor. Note the entire genome is not submitted as a predictor, as this would be nearly computationally impossible to solve. Covariates or even other predictors can be submitted to the model to ensure the effects observed are not attributable to other confounds. As described earlier, the effects observed in these models can be somewhat challenging to interpret in both a scientific and clinical setting. To make sense of these results, a structural equation model can be used to observe how the variance of the genetic architecture of each phenotype is explainable by their unique and shared properties.

The pipeline for a gSEM is a two-step process, with the first step being estimating a linkage disequilibrium score regression model. This model is quite specific to the genomics literature, and its computational details are dependent on concepts such as heritability. For our purposes, it was useful to conceptualize it as follows. Linkage disequilibrium (LD) can be thought of as the association between two sections of the genome that are not necessarily topographically related (i.e., one end of the genome to the other). Something to note is that these associations are not spurious, but rather meaningful variations caused by evolution. To get an LD score, we can correlate the j^{th} SNP with every other SNP in the genome to get the following score:



$$\hat{\beta}_j^{GWAS} = s_j + \sum_{k=1}^{J} \beta_k r_{x_{ij},x_{ik}} + e_j$$

s_j: bias from confounders

$r_{x_{ij},x_{ik}}$: correlation between SNPs x_j and x_k

e_j: estimation error

To model the shared and unique genomic pleiotropy of the disorder mentioned earlier, we first needed to curate a family of genomic summary statistics. Our study used summary statistics downloaded from the Psychiatric Genomics Consortium, UNC school of medicine. We intended to use the summary statistics to model different traits without individual SNP effects, run a common factor model, and conduct genetic multiple regressions. The summary statistics we chose were from models predicting PTSD, MDD, and SUD in large samples.

Methods

Genome-wide association studies (GWAS) is a useful technique for this project as it aims to understand the genetic causes of different human diseases. A GWAS is a study that identifies genetic variants in a population of interest vs. healthy control. For example, say we are curious what gene(s) might cause red hair, we can take all the DNA of a red-haired person, and compare it against the DNA of a person without red hair. The genes that are found to be different (after stringent statistical tests) are said to play a vital role in people having red hair. Furthermore, Structural equation models (SEMs) are modeling techniques commonly used in social and behavioral sciences that are built to handle multi-equation models, multiple measures of concepts, and measurement error SEM is more than one statistical technique (Bollen 2011). It incorporates different multivariate techniques into one model fitting framework (Sturgis, 2016). We chose to use this model because it is an integration of measurement theory, latent variable analysis, regression, path analysis, and simultaneous equations, most of which relate directly to our research question.

Methodology Formulations

First, three data files based on three traits have been downloaded from the Psychiatric Genomics Consortium, UNC school of medicine. The corresponding traits/mood disorders are major depressive disorder, substance use disorder, and post-traumatic stress disorder. All files already contained SNP id (type of genetic variation among people and each SNP represents a single DNA building block) (MedlinePlus, 2020), A1 allele (effect allele), and A2 allele (non-effect allele). Simple linear regressions were used to explore the downloaded data. However, linear relationships should be interpreted with caution because some variables may be just a transformation from another variable (A1 allele and OR, odds-ratio for the effect allele). Also, we are interested in the common factors and genetic correlations between different traits, and not concerned about the relationship between variables for an individual trait.

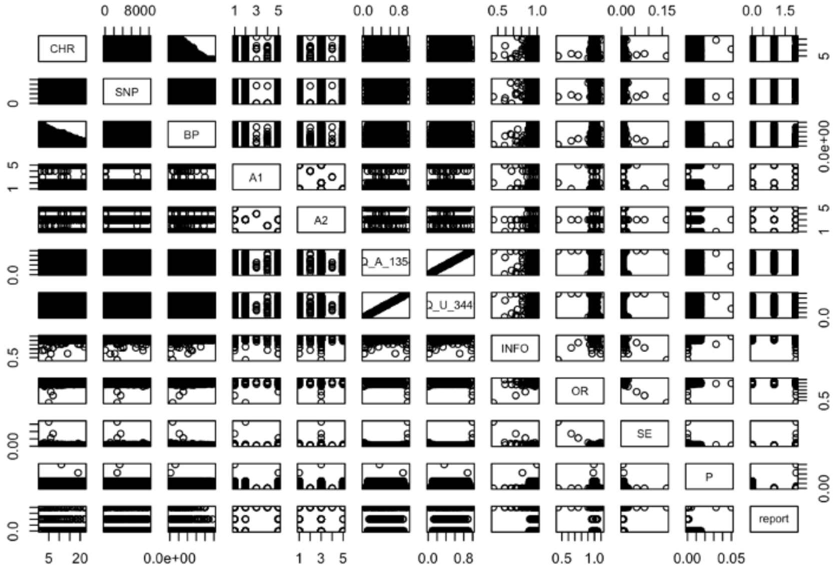


Figure 1: Exploratory Data Graphics for Major Depressive Disorder Raw File

Figure 1 depicts an exploration of the raw data file from the MDD study, this afforded us a better understanding of the variables. For instance a strong correlation between the genes in the middle of the matrix can be observed whereas the other relationships are a bit more random. After data exploration we “munged” trait files using the munge function from GenomicSEM library to generate the summary statistics, which will be used to calculate the covariance matrix. This essentially just cleaned out data so they were more organized and presented in a consistent way.

The second step is to calculate the genetic and sampling covariance matrix by running multivariable LD-Score regression. The LD (linkage disequilibrium) score regression can provide an accurate estimate of the genetic correlation between two traits. The regressions use the summary statistics for each trait, the sample prevalence vector (obtained from calculating cases/cases + controls), and the population prevalence vector (obtained from national health data).

The third step is to run models using the common factor model and specific structural models which provide insight into the shared genomic pleiotropy between these conditions. All models we run used diagonally weighted least square estimation.

Examples and Numerical Results

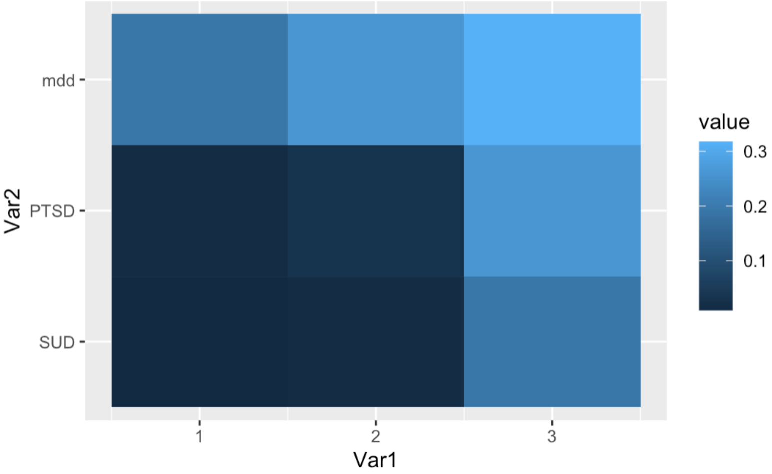
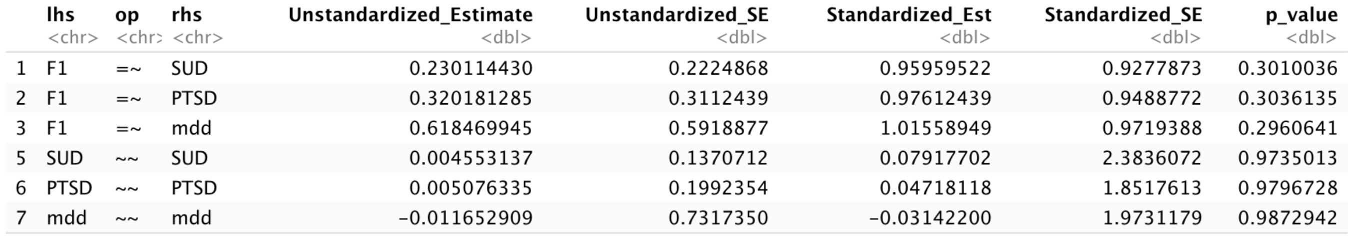


Figure 2: Genetic Covariance Matrix from running LD-score regression

Figure 2 demonstrates the covariance matrix on the liability scale for case/control designs. After running the common factor model, we found that the indicator loadings on the common factors are not significant for all three traits. Figure 3 represents the summary results of the common factor model.



	lhs	op	rhs	Unstandardized_Estimate	Unstandardized_SE	Standardized_Est	Standardized_SE	p_value
	<chr>	<chr>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	F1	=~	SUD	0.230114430	0.2224868	0.95959522	0.9277873	0.3010036
2	F1	=~	PTSD	0.320181285	0.3112439	0.97612439	0.9488772	0.3036135
3	F1	=~	mdd	0.618469945	0.5918877	1.01558949	0.9719388	0.2960641
5	SUD	~~	SUD	0.004553137	0.1370712	0.07917702	2.3836072	0.9735013
6	PTSD	~~	PTSD	0.005076335	0.1992354	0.04718118	1.8517613	0.9796728
7	mdd	~~	mdd	−0.011652909	0.7317350	−0.03142200	1.9731179	0.9872942

Figure 3:Common Factor Model Results

Other than the common factor model, we also ran three specific structural models to investigate the relationship between the three different traits.We constructed a model in which substance use disorder is regressed on major depressive disorder and post-traumatic stress disorder, a model in which post-traumatic stress disorder is regressed on major depressive disorder and substance use disorder, and a model in which major depressive disorder is regressed on substance use disorder and post-traumatic stress disorder. The results are summarized to produce the following path diagram, in which the significant relationships are highlighted:

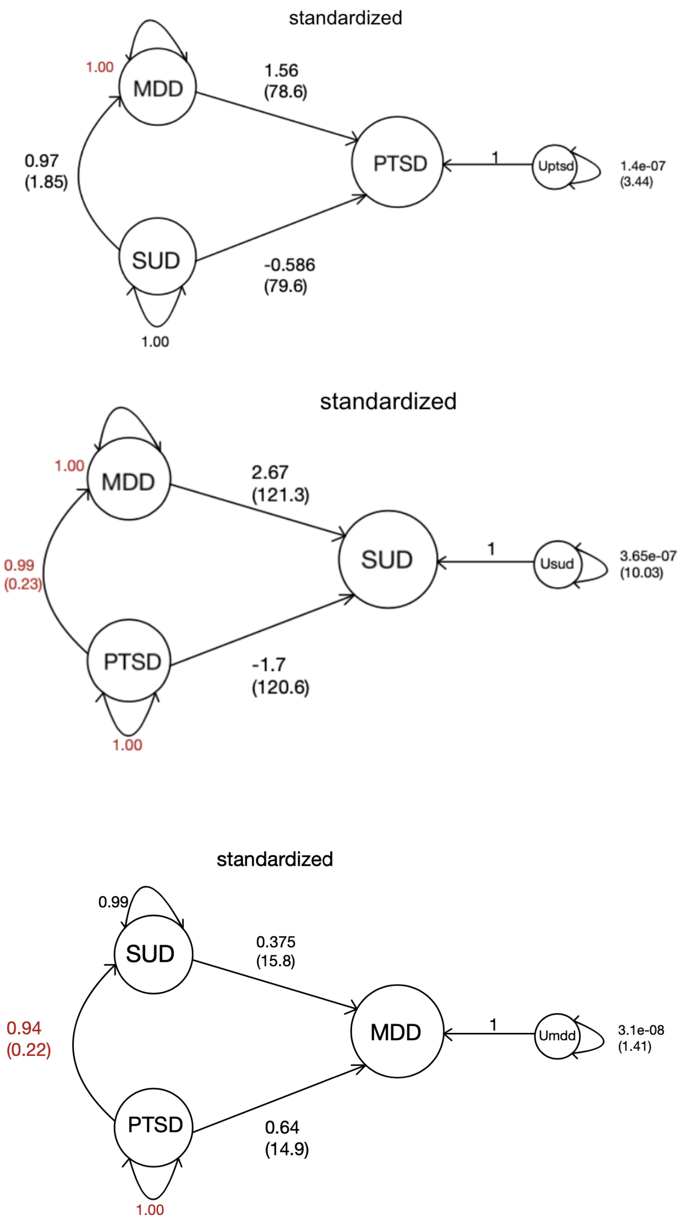


Figure 3: Interpreted Results in Path Diagrams, significant relationships are highlighted

(Note that some coefficients are greater than 1 after standardization, and it's referred to as the Heywood case. Even though it produces non-interpretable estimates of model fit, the values are not significant and shouldn't interpret too much from it).

Discussion and Conclusions

Our results found that the indicator loadings on the common factors are not significant in the common factor model, and the specific models didn’t find any significant association between predictor traits and outcome traits. Additionally, the correlation between SUD and PTSD in the standardized case is significant, and the correlation between MDD and PTSD in the standardized case is significant ($p<0.05$). The correlation between MDD and SUD is not significant ($p>0.05$). Our results suggest that even though some psychiatric disorders may be correlated, understanding the architecture of psychiatric illnesses requires more advanced evaluations and interpretations.

Additionally, SEM can offer some insight into what the exact nature of these correlations are . Without this novel technique we would likely keep running GWASs until there was no more data which would offer a heterogeneous and difficult to interpret picture. By leveraging the power of SEM to explain this variance in a way that allows us to determine the unique and shared genomic variants observed across disorders, we can begin to constrain our hypothesis space to particular sections of the genome. Once these areas of the genome are identified, we can then develop novel gene therapies that may one day prove to be imperative to the treatment of behavioral health disorders.

Works Cited:

Kennedy, D. N., Abraham, S. A., Bates, J. F., Crowley, A., Ghosh, S., Gillespie, T., Goncalves, M., Grethe, J. S., Halchenko, Y. O., Hanke, M., Haselgrove, C., Hodge, S. M., Jarecka, D., Kaczmarzyk, J., Keator, D. B., Meyer, K., Martone, M. E., Padhy, S., Poline, J. B., Preuss, N., … Travers, M. (2019). Everything Matters: The ReproNim Perspective on Reproducible Neuroimaging. *Frontiers in neuroinformatics*, *13*, 1. https://doi.org/10.3389/fninf.2019.00001

Manolio TA (July 2010). "Genomewide association studies and assessment of the risk of disease". *The New England Journal of Medicine*. **363** (2): 166–76. doi:10.1056/NEJMra0905980.

Bollen, K. A., & Noble, M. D. (2011). Structural equation models and the quantification of behavior. *Proceedings of the National Academy of Sciences*, *108*(Supplement_3), 15639–15646. https://doi.org/10.1073/pnas.1010661108

Grotzinger, A.D., Rhemtulla, M., de Vlaming, R. *et al.* Genomic structural equation modelling provides insights into the multivariate genetic architecture of complex traits. *Nat Hum Behav* 3, 513–525 (2019). https://doi.org/10.1038/s41562-019-0566-x

Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4. (2011). *Nature Genetics*, *43*(10), 977–983. https://doi.org/10.1038/ng.943

Youtube. (2014). *Professor Patrick Sturgis, Ncrm director, in the first (of three) part of the Structural Equation Modeling Ncrm online course*. *Youtube*.

Data & Code:

Download Results. Psychiatric Genomics Consortium. (2021, April 5). https://www.med.unc.edu/pgc/download-results/?wpv-category=major-depressive-disorder&wpv_aux_

current_post_id=4162&wpv_aux_parent_post_id=4162&wpv_sort_order=asc&wpv_view_count=4815

File format reference. File format reference - PLINK 1.9. (n.d.).

https://www.cog-genomics.org/plink2/formats