Utilizing Artificial Neural Networks to Identify Latent Network Pathways in Psychiatric Comorbidity

Georgia Smith

CSC/Math 4990, The University of Virginia's College at Wise
Advisor: Dr. James Vance

March 22, 2021

Contents

Chapter 1

Introduction

1.1 Overview

1.1.1 Introduction

Mental illness can be a burden for those whose lives are impacted from it, especially those with serious mental illness. This burden is exasperated when multiple mental illnesses co-occur - a phenomenon known as psychiatric comorbidity (further referred to as comorbidity).

Persons with serious mental illness are more likely to have comorbid medical conditions than those in the general population [12]. There also have been multiple studies identifying a high comorbidity between drug abuse and mental illness ([13], [11], [8]).

While the identification of illnesses commonly co-occuring is important for diagnosis and treatment, much is unknown about why diseases co-occur and what is causing them. The discovery of these causal factors and predictive features would have vast implications in the diagnosis and treatment of comorbid mental illness.

1.1.2 Latent versus Network Analysis

Recently, the field of psychometric analysis, or the analysis of psychological measurement techniques and theories, has had two major ideas of how two analyze comorbidity: a latent variable perspective and a network perspective. A latent variable perspective, as laid out by Boorsboom [2], is a combination of observed variables which allow inference of a latent variable.

Currently, the current Diagnostic Statistical Model of Mental Disorders, DSM-V, is considered a latent variable approach to mental illness. For diagnosis of a disorder (the latent variable) certain observable criteria must be met which show as symptoms. The latent variable, or disorder, is not just a label for when symptoms occur in the latent variable perspective. The latent variable is an unknown causal factor of the disorder, such as lack of serotonin and depression.

The network perspective does not discount the DSM-V, but instead does not focus on an unobservable causal factor. Instead, a diagnosis is caused and defined by an interactive *network* of observable symptoms. The network perspective was originally laid out by Cramer et. al in 2010 [3] using Alegria and Kessler's National Comorbidity Survey - Replication (NCS-R) [1].

Using the NCS-R, Kessler et. al. proposed a radical new way to conceptualize metal disorders and naturally leads to more complex network analysis. A main finding of Kessler et. al. was the discovery of "bridge symptoms," which are symptoms that occur in multiple disorders connecting them.

1.2 Goal

We hope to extend Cramer et. al.'s [3] initial hypothesis of the usability of network analysis to latent network path analysis, identifying pathways of comorbidity to identify causal symptoms of disorders. We will do this in a multi-stage production.

First, we will develop networks of individuals in the NCS-R building directed graphs based on the onset age of symptoms. Then, we will combine all of the networks for individuals with a certain diagnosis. After generating graphs for all of the diagnoses in Table B.1 we will identify cases of comorbidity and cobine the networks of individuals with comorbidity. Finally, we will compare the comorbidity networks with the diagnosis networks to identify bridge symptoms.

1.3 Data

We will utilize the data from the NCS-R [1]. The NCS-R is comprised of 3,713 variables which represent questions from a questionnaire designed to aid in diagnosis based on the DSM-V. The questionnaire is split into 46 sections with questions specific to specific disorders - such as depression, post-traumatic stress disorder (PTSD), and dementia - as well as demographics and lifetime events.

The NCS-R is a representative survey of adults (age 18+) in the United States,

comprised of 9,282 subjects. The NCS-R was completed in the homes of participants and all participants that met criteria for a mental disorder were issued a second, more in-depth interview [7].

The demographics of the NCS-R can be seen in Tables 1.1 - 1.4. All of the DSM-IV Diagnosis can be seen in Table B.1 in the appendix.

Age	**18-29**	**30-44**	**45-59**	**60+**
Percent	22.7	31.7	24.6	21.0

Table 1.1: Age Distribution of the NCS-R [7]

Sex	**M**ale	**F**emale
Percent	44.6	55.4

Table 1.2: Sex Distribution of the NCS-R [7]

Race	**N**on-**H**ispanic **W**hite	**N**on-**H**ispanic **B**lack	**H**ispanic	**O**ther
Percent	72.1	13.3	9.5	5.1

Table 1.3: Race Distribution of the NCS-R [7]

Region	**N**ortheast	**M**idwest	**S**outh	**W**est
Percent	18.4	26.7	34.5	20.5

Table 1.4: Regional Distribution of the NCS-R [7]

1.4 Contribution

While a network analysis approach to comorbidity has been investigated before, minimal work has been done to identify the direction of the edges between symptoms to identify causal symptoms. Main contributions to the field include:

- Latent Network Path Analysis on Comorbidity Data
- Neural Networks used in a Network Approach
- Identification of Causal Symptoms using Network Analysis
- Extension on Bridge Symptom Identification

Chapter 2

Literature Review

2.1 Cramer et. al., 2010 [3]

Cramer et. al. sought to rethink the idea of causal relationships in psychiatry in their 2010 paper. They advocated for a network perspective that says "disorders are *networks* that consist of *symptoms* and *causal* relations between them." [3]

A cornerstone of Cramer et. al.'s argument is an assumption that the latent variable model cannot include the cyclic networks which support the causal relationships between symptoms (i.e. you're anxious and trying not to be which makes you more anxious) [3]. Danks et. al. disagreed with this assumption, arguing instead that we can define the latent variable model to do this and assuming inability to do so limits possibilities [4].

In this paper we will propose a latent-network hybrid where a network model is used to identify latent, directed pathways in comorbidity.

Cramer et. al. focused on the comorbid (or co-occuring) relationship between Major Depression Disorder (MDD) and Genral Anxiety Disorder (GAD). The began with the theory of complex networks "without assuming *a priori* that scuh relationships arise from a mental disorder as a common cause" [3]. Then, Cramer et. al. put symptoms into nodes and created paths to represent the relation between symptoms.

They used statistical parameterization and the Akaike Information Criteron to find the most accurate model. They found that a bridge model holds when there are no independent variables. A bridge model is a undirected graph where overlapping nodes (symptoms) indicate a comorbid relationship (see Figure 2.1) [3].

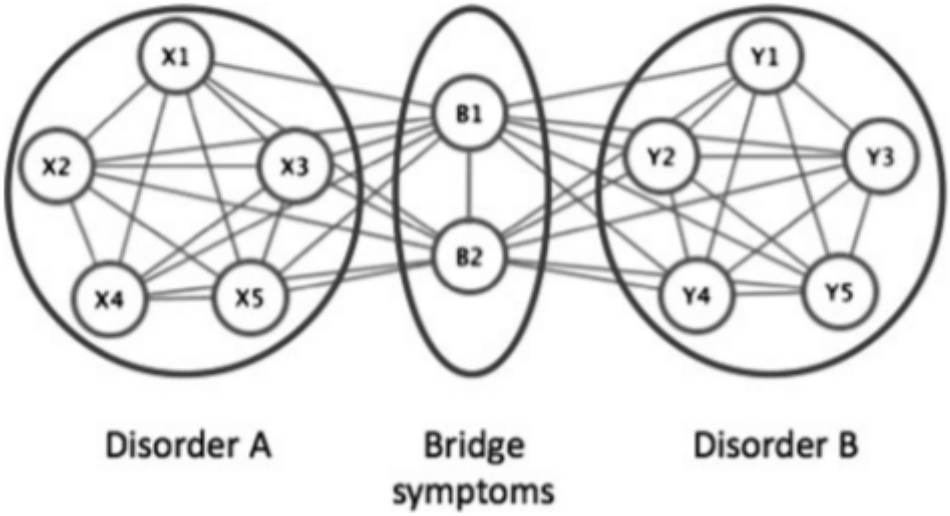

Figure 2.1: Example of a Bridged Network [3]

Using this bridge model, Cramer et. al. used edge thickness and color to further demonstrate relationships between symptoms. The edge thickness represented the co-occurrence of the two symptoms and the edge color represented the strength or association, or log odds ratio, between the two symptoms. The node size was used to exemplify frequency while the color is the node strength, or the sum of the weights of the connected edges [3]. The outcome of these additions can be seen in Figure 2.2.

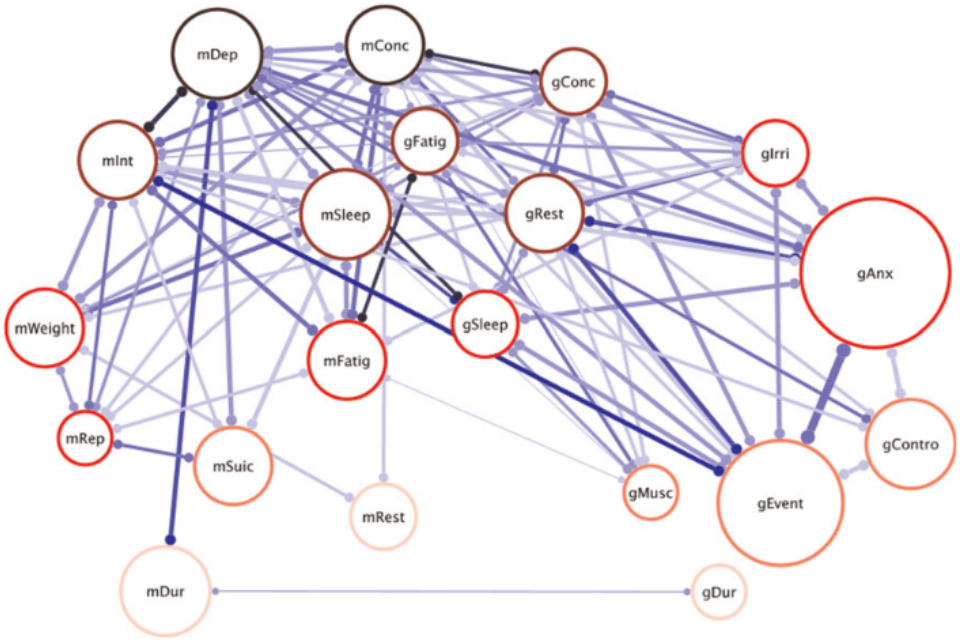

Figure 2.2: Example of a Fully Realized Network using GAD and MDD [3]

Cramer et. al. found that a network approach can be very useful. First, it can aid in hypothesis development on the cause of psychiatric disorders and specific symptoms. Secondly, the can identify "pathways of comorbidity" that can help lead to identifying some root cause of symptoms (i.e. MDD turning into GAD or *vice versa*). They also found that symptoms were more strongly connected when there was at least one pair of overlapping symptoms [3].

Overall, Cramer et. al. found that a network approach to psychometric analysis is a potentially groundbreaking approach to the analysis of comorbidity and can lead to discovery about the causal relationship between symptoms [3].

2.2 Qian et. al. 2020 [10]

Qian et. al. published a paper in 2020 identifying latent pathways in comorbidity using a dataset comprised of colorectal cancer patients. Qian et. al. built on current point process models to develop a new process - Deep Diffusion Process (DDP) [10].

DDP was realized using a continuous-time recurrent neural network (RNN)

using sigmoid normalization and long short term memory. These methods allow for not only the identification of network pathways, but also judge the probability of future event given a different event has happened [10].

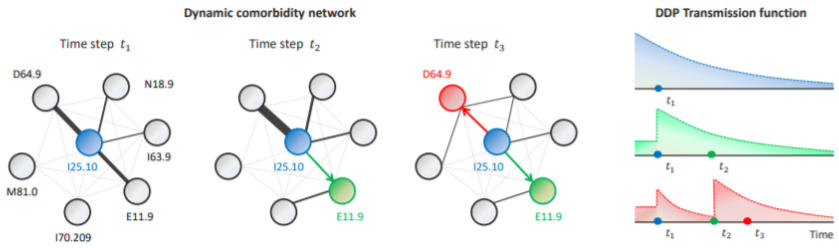

Figure 2.3: Qian et. al.'s realization of DDP. Nodes are identified by their ICD-10 code [10]

Figure 2.3 shows the impace DDP can have. The onset of heart disease is at t_1, which increases the possibility of anemia, diabetes, renal failure, and cerebral infarction. When diabetes onset begins at timestep t_2 that increases the probability of the ilnesses from t_1, and creates a connection to atherosclerosis.

This method is not revolutionary, but in testing, Qian et. al.'s DDP model outpreformed all other point process models with the comorbidity data [10].

Chapter 3

System Analysis and Design

3.1 Implementation Platform

We used multiple platforms in the development of this project. They code is written in Python version 3.8.5. Multiple IDEs were used including Visual Studio Code, VIM, and Jupyter Notebook. We also utilized University of Virginia's Rivanna High Performance Computing (HPC) cluster and acknowledge Research Computing at The University of Virginia for providing computational resources and technical support that have contributed to the results reported within this publication (https://rc.virginia.edu).

3.1.1 Programming Languages

Python was chosen as the primary programming language of this project due to its touted versatility [9], our prior expertise in the language, and because it Python is the language used in multiple prior works on latent networks providing us libraries and modules to utilize.

3.1.2 Description of Programming Environment

We first developed code in Jupyter Notebook for testing and debugging. Once the code was working satisfactorily we merged it into a Python file using Visual Studio Code. This code was then either run on a local machine or exported to the Rivanna HPC cluster, depending on computational intensity of the code. We then took code output, in the form of a CSV or serialized with Pickle, and imported it

into Jupyter Notebook for further analysis and visualization develpment. Visualizations were developed using the NetworkX [5] and MatPlotLib [6] Libraries.

3.2 Design Overview

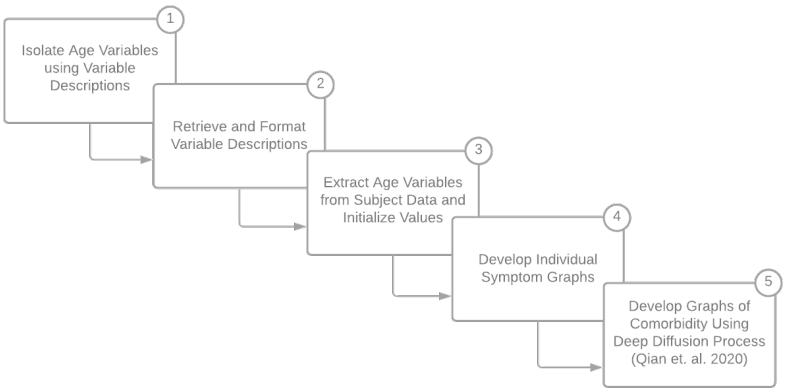

Figure 3.1: System Design Overview

We use a multi-staged process for our system outlined in Figure 3.1. We begin by processing the data to initialize the values and extract relevant information. Then, we develop graphs to describe individual subjects. We finish by implementing Deep Diffusion Process (DDP) [10].

3.3 Input

The code we developed only requires Kessler's NCS-R dataset [1]. For comorbidity analysis two DSM-IV diagnosis listed in Table B.1 need to be selected. Analyses can also be run on single diagnosis instead of comorbid by selecting on diagnosis from Table B.1.

3.4 Data Preprocessing

3.4.1 Data Descriptions

The first step in our software design is to get the descriptions for the variables in the NCS-R dataset. As described in section 1.3 there are 3,714 variables in the NCS-R. The NCS-R is hosted on the Inter-university Consortium for Political and Social Research (ICPSR) at University of Michigan. ICPSR includes a section using the Survey Documentation and Analysis (SDA) package which allows users greater insight into the data including descriptions for each variable. This information can also be accessed as a JavaScript array containing all variables and their description or queried by variable to access JSON objects.

Figure 3.2: Process Used to Build the NCS-R Data Description Tree

The JavaScript array describing variables in the NCS-R dataset is split into 43 sections mirroring the questionnaire given to subjects. This array is accessible at https://www.icpsr.umich.edu/SDA/NAHDAP/20240-0002/CODEBOOK/tree_items.js when logged in with an ICPSR account.

We used the JavaScript array to build a data tree so that we could easily access data descriptions. The process used to build the tree is outlined in Figure 3.2.

The original JavaScript array includes extraneous JavaScript data and combines the variable name and description into a single string. We applied the regular expression *["([\V]+) - ([\V]+)", "j[\V+"],* which was used to search and replace with *["$1", "$2"]*, resulting in the format *["Variable Name", "Variable Description"].*

We also scanned through the array multiple times and checked for errors due to patterns not matching the regular expression above. Once the array was formatted we iterated over the JavaScript array recursively to build the tree.

3.4.2 Age Variable Subset

The NCS-R questionnaire is very thorough and includes a large amount of boolean questions. We needed to filter the variables representing those questions out so we could isolate variables having to do with age (further referred to age variables) that we could use as time series data. We analyzed the description data outlined above and found patterns in the descriptions for age variables.

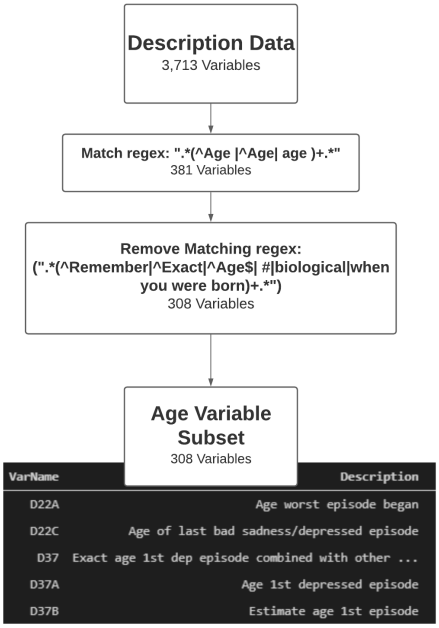

Figure 3.3: Process Used to Extract Age-Based Variables from the NCS-R Data Descriptions

With the patterns isolated we developed regular expressions to extract the age variable subset of the NCS-R. We found that doing this in a two-tiered approach worked best. This approach is outlined in Figure 3.3.

3.4.3 Data Initialization

With the age variables subset we could then isolate them in the NCS-R Survey Response Dataset (further referred to as just NCS-R or NCS-R Data).

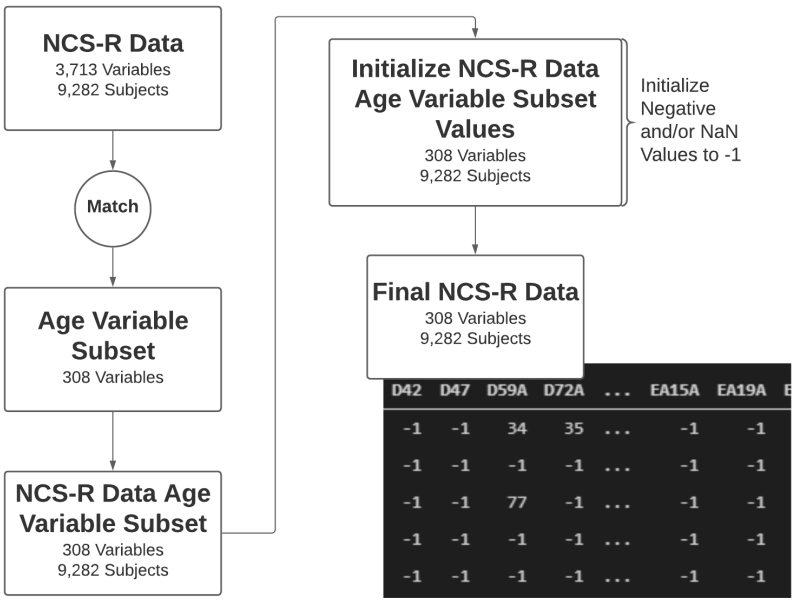

Figure 3.4: Process Used to Initializes Values in the NCS-R Survey Response Dataset

Figure 3.4 outlines this process. In the NCS-R subject have options to not respond and some questions are not applicable to all subjects. These values are blank, input as *Not a Number* (NaN), or set as negative values. We initialized all of those values to -1 for ease of analysis.

3.5 Deep Diffusion Process

With the data subset to values that could be used as time series we used Qian et. al.'s Deep Diffusion Process (DDP) to obtain weighted edge values for graphs [10]. We preformed this on all subjects, subjects subset by DSM-IV Diagnosis, and subjects subset by comorbid diagnosis.

DDP takes data by subject as a list of dictionaries with the values *idx_event*, *type_event*, *time_since_start*, and *time_since_last_event*. We used Algorithm 1 to alter the NCS-R data to a format usable with DDP.

Algorithm 1 NCS-R Data to DDP Format

for $subject = 1, 2, \ldots$ **do**
    Sort $subject$ lifetime events by age
    for $event = 1, 2, \ldots$ **do**
        $event.idx\_event = Event$
        $event.type\_event = Event\ Description$
        $event.time\_since\_start = Age\ at\ Event$
        $event.time\_since\_last\_event = Last\ Event\ Age - Age\ at\ Event$
    end for
end for

Bibliography

[1] Alegria, M., Jackson, J. S. J. S., Kessler, R. C., and Takeuchi, D. (2016). Collaborative psychiatric epidemiology surveys (cpes), 2001-2003 [united states]. 3, 11

[2] Borsboom, D. (2008). Latent variable theory. 2

[3] Cramer, A. O., Waldorp, L. J., Van Der Maas, H. L., and Borsboom, D. (2010). Comorbidity: A network perspective. *Behavioral and brain sciences*, 33(2-3):137. 0, 3, 6, 7, 8

[4] Danks, D., Fancsali, S., Glymour, C., and Scheines, R. (2010). Comorbid science? *Behavioral and Brain Sciences*, 33(2-3):153–155. 6

[5] Hagberg, A. A., Schult, D. A., and Swart, P. J. (2008). Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T., and Millman, J., editors, *Proceedings of the 7th Python in Science Conference*, pages 11 – 15, Pasadena, CA USA. 11

[6] Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95. 11

[7] Kessler, R. C., Berglund, P., Chiu, W. T., Demler, O., Heeringa, S., Hiripi, E., Jin, R., Pennell, B.-E., Walters, E. E., Zaslavsky, A., et al. (2004). The us national comorbidity survey replication (ncs-r): design and field procedures. *International journal of methods in psychiatric research*, 13(2):69–92. 4, 19, 20, 21, 22, 23

[8] Kokkevi, A. and Stefanis, C. (1995). Drug abuse and psychiatric comorbidity. *Comprehensive Psychiatry*, 36(5):329–337. 2

[9] Prendergast, L. (2019). What makes python so versatile? *Medium*. 10

[10] Qian, Z., Alaa, A., Bellot, A., Schaar, M., and Rashbass, J. (2020). Learning dynamic and personalized comorbidity networks from event data using deep diffusion processes. In *International Conference on Artificial Intelligence and Statistics*, pages 3295–3305. PMLR. 0, 8, 9, 11, 15

[11] Regier, D. A., Farmer, M. E., Rae, D. S., Locke, B. Z., Keith, S. J., Judd, L. L., and Goodwin, F. K. (1990). Comorbidity of mental disorders with alcohol and other drug abuse: results from the epidemiologic catchment area (eca) study. *Jama*, 264(19):2511–2518. 2

[12] Sokal, J., Messias, E., Dickerson, F. B., Kreyenbuhl, J., Brown, C. H., Goldberg, R. W., and Dixon, L. B. (2004). Comorbidity of medical illnesses among adults with serious mental illness who are receiving community psychiatric services. *The Journal of nervous and mental disease*, 192(6):421–427. 2

[13] Volkow, N. D. (2001). Drug abuse and mental illness: progress in understanding comorbidity. *American Journal of Psychiatry*, 158(8):1181–1183. 2

Appendix A

Figure Appendix

Appendix B

Table Appendix

B.1 NCS-R Data Tables

Table B.1: DSM Diagnosis Counts in the NCSR [7].

Begin of Table	
DSM Diagnosis	**Count**
DSM-IV Attention Deficit Disorder (LifeT)	365
DSM-IV Agoraphobia without Panic Disorder (LifeT)	231
DSM-IV Agoraphobia with Panic Disorder (LifeT)	126
DSM-IV Alcohol Abuse (Lifetime)	1034
Lifetime Alcohol Abuse w/ hierarchy	590
DSM-IV Alcohol Dependence (Lifetime)	444
DSM-IV Adult Separation Anxiety Disorder (LifeT)	558
DSM-IV Bi-Polar I (Lifetime)	101
DSM-IV Bi-Polar II (Lifetime)	105
Lifetime Bi-Polar Subthreshold	210
DSM-IV Conduct Disorder (Lifetime)	405
DSM-IV Drug Abuse (Lifetime)	651
DSM-IV Drug Dependence (Lifetime)	248

Continuation of Table B.1 - DSM Diagnosis Counts in the NCSR [7]

DSM Diagnosis	**Count**
DSM-IV Dysthymia (Lifetime)	386
DSM-IV Dysthymia with hierarchy (LifeT)	232
DSM-IV Generalized Anxiety Disorder (LifeT)	752
DSM-IV Gen Anxiety Disorder w/hierarchy (LifeT)	553
DSM-IV Hypomania (Lifetime)	77
DSM-IV Intermittent Explosive Disorder (LifeT)	728
DSM-IV Intermittent Explosive Disorder w/ hierarchy (LifeT)	678
DSM-IV Mania (Lifetime)	339
DSM-IV Major Depressive Disorder w/ hierarchy (LifeT)	1579
DSM-IV Major Depressive Episode (Lifetime)	1829
DSM-IV Oppositional Defiant Disorder (LifeT)	453
DSM-IV Oppositionl Defiant Disorder w/ hierarchy (LifeT)	375
DSM-IV Panic Attack (Lifetime)	2573
DSM-IV Panic Disorder (Lifetime)	455
DSM-IV Posttraumatic Stress Disorder (LifeT)	604
DSM-IV Separation Anxiety Disorder (LifeT)	331
DSM-IV Social Phobia (Lifetime)	1143
DSM-IV Specific Phobia (Lifetime)	1198
DSM-IV Nicotine Dependence (Lifetime)	626
DSM-IV Alcohol Abuse (30Day)	64
DSM-IV Alcohol Abuse w/hierarchy (30Day)	31
DSM-IV Alcohol Dependence (30 day)	43
DSM-IV Adult Separation Anxiety Disorder (30Day)	68
DSM-IV Drug Abuse (30 day)	32
DSM-IV Drug Abuse w/ hierarchy (30 day)	18

Continuation of Table B.1 - DSM Diagnosis Counts in the NCSR [7]

DSM Diagnosis	**Count**
DSM-IV Drug Dependence (30 day)	14
DSM-IV Dysthymia (30 day)	122
DSM-IV Dysthymia with hierarchy (30 day)	71
DSM-IV Generalized Anxiety Disorder (30Day)	157
DSM-IV Gen Anxiety Disorder w/hierarchy (30Day)	97
DSM-IV Hypomania (30 day)	9
DSM-IV Intermittent Explosive Disorder (30day)	161
DSM-IV Intermittent Explosive Disorder w/ hierarchy (30Day)	151
DSM-IV Mania (30 day)	65
DSM-IV Major Depressive Disorder w/ hierarchy (30Day)	233
DSM-IV Major Depressive Episode (30 day)	301
DSM-IV Panic Attack (30 day)	306
DSM-IV Panic Disorder (30 day)	105
DSM-IV Social Phobia (30 day)	334
DSM-IV Specific Phobia (30 day)	592
DSM-IV Agoraphobia without Panic Disorder (30Day)	77
DSM-IV Agoraphobia with Panic Disorder (30Day)	38
DSM-IV Nicotine Dependence (30 day)	193
DSM-IV Anorexia (Lifetime)	21
DSM-IV Binge Eating Disorder w/ hierarchy (Lifetime)	105
DSM-IV Binge Any (Lifetime)	192
DSM-IV Bulimia (Lifetime)	53
DSM-IV Bulimia w/ hierarchy (Lifetime)	52
DSM-IV Binge Eating Disorder w/ hierarchy (12Mo)	51
DSM-IV Binge Any (12Mo)	86

Continuation of Table B.1 - DSM Diagnosis Counts in the NCSR [7]

DSM Diagnosis	**Count**
DSM-IV Bulimia (12Mo)	16
DSM-IV Bulimia w/ hierarchy (12Mo)	16
DSM-IV Attention Deficit Disorder (12Mo)	190
DSM-IV Agoraphobia without Panic Disorder (12Mo)	138
DSM-IV Agoraphobia with Panic Disorder (12Mo)	73
DSM-IV Alcohol Abuse (12Mo)	213
DSM-IV Alcohol Abuse w/ hierarchy (12Mo)	113
DSM-IV Alcohol Dependence (12 month)	106
DSM-IV Adult Separation Anxiety Disorder (12Mo)	156
DSM-IV Bi-polar I (12Mo)	65
DSM-IV Bi-polar II (12Mo)	74
DSM-IV Bi-Polar Subthreshold (12Mo)	123
DSM-IV Conduct Disorder (12 month)	33
DSM-IV Drug Abuse (12 month)	102
DSM-IV Drug Abuse w/ hierarchy (12 month)	61
DSM-IV Drug Dependence (12 month)	36
DSM-IV Dysthymia (12 month)	226
DSM-IV Dysthymia w/hierarchy (12 month)	137
DSM-IV Generalized Anxiety Disorder (12Mo)	394
DSM-IV Gen Anxiety Disorder w/hierarchy (12Mo)	261
DSM-IV Hypomania (12 month)	32
DSM-IV Intermittent Explosive Disorder (12Mo)	404
DSM-IV Intermittent Explosive Disorder w/ hierarchy (12Mo)	377
DSM-IV Mania (12 month)	190
DSM-IV Major Depressive Disorder w/ hierarchy (12Mo)	658

Continuation of Table B.1 - DSM Diagnosis Counts in the NCSR [7]	
DSM Diagnosis	**Count**
DSM-IV Major Depressive Episode (12Mo)	805
DSM-IV Oppositional Defiant Disorder (12Mo)	55
DSM-IV Oppositionl Defiant Disorder w/ hierarchy (12Mo)	48
DSM-IV Panic Attack (12 month)	995
DSM-IV Panic Disorder (12 month)	262
DSM-IV Posttraumatic Stress Disorder (12Mo)	326
DSM-IV Social Phobia (12 month)	652
DSM-IV Specific Phobia (12 month)	843
DSM-IV Nicotine Dependence (12 month)	312
DSM-IV Binge Eating Disorder w/ hierarchy (30Day)	28
DSM-IV Binge Any (30Day)	45
DSM-IV Bulimia (30Day)	8
DSM-IV Bulimia w/ hierarchy (30Day)	8
DSM-IV Bi-polar I (30Day)	35
DSM-IV Bi-polar II (30Day)	37
DSM-IV Bi-Polar Subthreshold (30Day)	41
DSM-IV Posttraumatic Stress Disorder(30Day)	160
End of Table B.1 - DSM Diagnosis Counts in the NCSR [7]	

Appendix C

Pseudo-Code Appendix