Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Praveen Suthaharan Date

Latent Class Analysis for PTSD Subtype Discovery

By

Praveen Suthaharan

Master of Science in Public Health

Department of Biostatistics and Bioinformatics

_____

Suprateek Kundu, Ph.D.

Committee Chair

_____

Benjamin Risk, Ph.D.

Committee Member

Latent Class Analysis for PTSD Subtype Discovery

By

Praveen Suthaharan

B.S. (double major)
North Carolina State University
2016

Thesis Committee Chair: Suprateek Kundu, Ph.D.

An abstract of
A thesis submitted to the Faculty of the
Rollins School of Public Health of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Biostatistics
2019

Abstract

Latent Class Analysis for PTSD Subtype Discovery

By Praveen Suthaharan

Background: Exposure to trauma presents a public health concern worldwide. Common traumatic events include child abuse, military combat, personal assaults, and car accidents. It has been shown that more than half of the female and male population experience at least one traumatic event in their lifetime. Although more than half experience trauma, only a small percentage develop post-traumatic stress disorder (PTSD). Moreover, little is understood about the underlying neurobiological mechanism contributing to this extreme heterogeneity in PTSD.

Objective: To discover PTSD subtypes in our cohort study for explaining the extreme symptomatologic heterogeneity.

Methods: A cluster analysis was performed on the Grady Trauma Project study. The cluster analysis involved a Dirichlet Process (DP)-based Latent Class Analysis (LCA) to discover PTSD subtypes. We performed a non-parametric Bayesian technique, DP, in conjunction with the LCA to non-empirically discover subtypes of PTSD.

Results: The clustering analysis revealed 4 distinct subtypes of patients with resulting groups of 23 patients, 15 patients, 9 patients and 31 patients in each of the groups, respectively. Likewise, the three clinically-defined symptom (intrusive, avoidance/numbnesss, and hypearousal) categories characterize the PTSD subtypes into clear, separable clusters – cluster 1 with *moderate- to high- intrusive symptom-present* patients, *low- to high- avoidance/numbness symptom-present* patients, *low- hyperarousal* patients, cluster 2 with *moderate- to high- symptom-present* patients, cluster 3 with *high symptom-present* patients, and cluster 4 with *symptom-absent to low symptom-present* patients.

Conclusions: Our research reveals discovery of PTSD subtypes as a benchmark for explaining the inherent symptomatologic heterogeneity in PTSD symptom profiles. However, this raises important questions regarding the association between the underlying neurobiological mechanism and behavioral difference. We aim to further explain the symptomatologic heterogeneity through brain network analyses to discover important brain connectivity patterns that influence the onset of the various symptoms of PTSD.

Latent Class Analysis for PTSD Subtype Discovery

By

Praveen Suthaharan

B.S. (double major)

North Carolina State University

2016

Thesis Committee Chair: Suprateek Kundu, Ph.D.

A thesis submitted to the Faculty of the

Rollins School of Public Health of Emory University

in partial fulfillment of the requirements for the degree of

Master of Science in Public Health

in Biostatistics

2019

Acknowledgements

I would like to express my sincere gratitude to the Department of Biostatistics and Bioinformatics at Emory for promoting high-quality biostatistics education and research to all students and faculty. Undoubtedly, I am deeply indebted to Dr. Suprateek Kundu. Without his guidance and support, and insight in the field of neuroimaging research, I would not be able to embrace my current determination and passion to pursue a future career as a researcher in computational psychiatry.

The establishment of my thesis is largely attributed to the provision of both the clinical and brain imaging data from The Grady Trauma Project. I would like to thank both our collaborators – Dr. Jennifer Stevens, for explaining the clinical data, and Dr. Negar Fani for facilitating the brain imaging analysis. Moreover, I would like to offer my appreciation to three Ph.D. students – Josh Lukemire, Zae Higgins, and Xin Ma – for supporting me through all the adversities encountered in my thesis work and Dr. Benjamin Risk for providing important feedback on my thesis.

I would also like to show my appreciation to the faculty, students and staff that make up the closely-knit biostatistics family. Their undying moral support promoted a comfortable and positive stress-free environment.

Most importantly, I would like to thank my friends and family for their continuous encouragement. My father, Dr. Shanmugathasan Suthaharan, for his knowledge and guidance which helped me overcome much of my adversity in compiling my thesis as well as his fastidious proofreading. My mother, Manimehala Suthaharan, for her loving support and care for my well-being. Lastly, my sisters – Lovepriya Suthaharan and Prattheeba Suthaharan – for their constant support through proofreading of my thesis as well.

Table of Contents

Introduction

Trauma exposure presents a global public health concern. Some of the most common traumatic exposures include child abuse, military combat, and car accidents. The National Comorbidity Survey (NCS) reports that 60.7% of men and 51.2% of women experience at least one traumatic event, with a significant proportion of these events occurring during childhood.[1] Among those who experience traumatic events, only 10-40% develop psychiatric symptoms of clinical relevance,[2-5] such as those found in post-traumatic stress disorder (PTSD).

The Diagnostic and Statistical Manual (DSM) is typically used as the diagnostic criteria for mental disorders, including PTSD.[6] After a certain period of traumatic exposure, based on the DSM, symptoms of PTSD emerge into PTSD. PTSD is primarily defined by three main clinically-relevant symptom clusters: (1) re-experiencing (i.e., intrusive), (2) avoidance and emotional numbing, and (3) hyperarousal, although the latest DSM (DSM-V) further divides the second cluster resulting in four clusters. Each of these symptom clusters is further stratified into multiple symptoms based on the 17 question PTSD Symptom Scale (PSS). On this scale, symptoms 1-4,17 are of intrusive-type, symptoms 5-11 are of avoidance and numbness-type, and symptoms 12-16 are of hyperarousal-type. This use of multiple defining symptoms raises the problem of heterogeneity.

Despite efforts to understand the psychopathology of PTSD symptoms, there still remain important unanswered questions regarding symptomatologic differences of patients and whether certain symptoms of PTSD play important roles in the brain network topology.

One of the challenges in addressing our question of interest is the extreme heterogeneity seen in the symptom profiles of PTSD patients. Studies have shown that over half of the US population experience some form of trauma in their lives, but only 7-8% show symptoms of PTSD.[1,7] Although factors like personal/family history of psychiatric disorders, early lifetime traumatic experience, and/or gender (i.e., predominantly female) were shown to be associated with elevated risks for developing PTSD, they may not fully explain the heterogeneous symptom characteristic of PTSD.[2,8]

Current neurobiological models postulate that PTSD symptoms rise from dysfunctions in fear extinction caused by traumatic events.[9] Fear extinction is our neurobiological ability to adapt, as situations change, by learning to suppress a previously learned fear. In individuals with PTSD, this suppressed fear comes back and cannot be completely erased. This suggests that deficits in fear extinction retention underlies PTSD, leading to the persistence of PTSD-related symptoms.[10] Many brain regions have been shown to contribute to the onset of anxiety disorders.[11] However, no single brain region is known to be strictly responsible for causing anxiety disorders. Therefore, it is of increasing interest to identify dysfunctions in brain circuits that yield PTSD-related behavioral symptoms. These dysfunctional brain circuits are usually identified based on differences in functional connectivity across brain regions between healthy individuals and individuals with mental conditions of interest.[12] Existing brain imaging studies have been conducted to classify individuals with different PTSD status; however, these studies do not focus on investigating significant brain circuits that drive symptomatologic differences of PTSD patients.[13,14] There is also a need for prediction and inferential tools that utilize symptoms of PTSD and the brain network to identify significant network features that drive

symptomatologic differences. Therefore, it is of importance to investigate brain circuits in hopes of revealing significant neurobiological biomarkers for optimal diagnosis and treatment of PTSD. It is obvious that trauma-induced PTSD is a significant public health concern and addressing the underlying biological mechanisms driving the heterogeneity of the psychopathology of PTSD symptoms is of paramount importance.

With the advent of robust neuroimaging technology, there are unprecedented opportunities to investigate how traumatic events interact with brain circuits which lead to PTSD-related behavioral symptoms. There are several major challenges in achieving this goal:

i. the brain imaging data is high dimensional and noisy
ii. it is unclear how to identify relevant brain connectome features as biological phenotypes for PTSD-related symptoms

modeling relationship between the high dimensional brain network with behavioral outcomes is complex.

In this thesis, we investigated PTSD subtypes to explain the symptomatologic heterogeneity issue through subtype analysis and propose a network analysis to investigate whether connectivity among certain brain regions were predictive of symptoms of PTSD. We focused our examination at the level of the symptom as illustrated in Figure 1 (Stage II Analysis is described in more detail in the *Further Research* section).

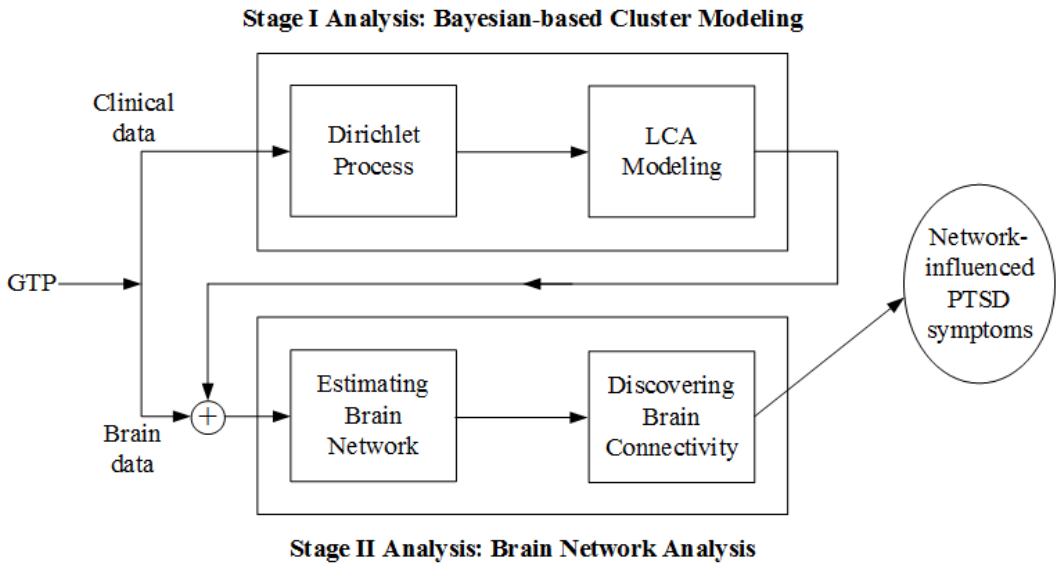


Figure 1. Process diagram of two-stage approach. The two-stage approach involves analysis on both the clinical data and brain data taken from the Grady Trauma Project (GTP). In Stage I, we cluster the data based on clinical features using the Dirichlet Process (DP) and Latent Class Analysis (LCA). In Stage II (**discussed in the *Further Research***), using the cluster output from Stage I, we perform estimation of the brain network to discover brain connectivity patterns in order to yield evidence for network-influenced PTSD symptoms.

First, we used a non-parametric Bayesian-based latent class analysis (LCA) to identify subtypes (or clusters) of patients sharing commonalities among symptoms. In a prior study investigating subtypes of PTSD symptoms, four subclasses were suggested: *high severity and comorbidity* symptoms, *moderate severity* symptoms, *low* symptoms, and a *resilient* class with low symptoms.[15] We were more interested in performing cluster analysis on the basis of a Bayesian framework to overcome the process of empirically selecting the number of clusters.

Second, as discussed in the *Further Research*, we will introduce network analyses using a Bayesian binary logistic regression model to map functional connectivity patterns to each symptom in order to discover important connections that are predictive of the symptoms

of PTSD. Current research regarding the effects of PTSD symptoms on the network topology have shown significant associations of emerging between symptom pairs as evident by strong network connection using correlation models.[16,17] However, there are limited applications of Bayesian modeling for functional connectivity inference at the level of the symptoms

The rest of this thesis is organized as follows. Section 1 describes the methodology of the clustering. Section 2 presents subgroup identification and characterization from the clustering technique. Section 3 summarizes the results found from the clustering analysis and discusses future extension in the setting of network analysis. Section 4 emphasizes certain limitations of the analysis and presents the conclusions.

Methods

We implement a systematic approach for discovering PTSD subtypes. This approach provides a framework for revealing hidden groups in a flexible manner. The Bayesian-based framework is illustrated in Figure 2. We use a non-parametric Bayesian technique in conjunction with a finite mixture model to discover PTSD subtypes.

Data description

Grady Trauma Project (GTP)

We use data from the Grady Trauma Project (GTP), a publicly funded, tertiary care center serving a predominantly socioeconomically disadvantaged inner-city population. The GTP cohort study recruited participants with a history of civilian trauma exposure for the last 12 years and includes data on childhood and adult trauma, as well as PTSD (PTSD symptom

scale, or PSS). Subjects (n=78) used in this analysis were all females, primarily African American, low socioeconomic and inner-city population.

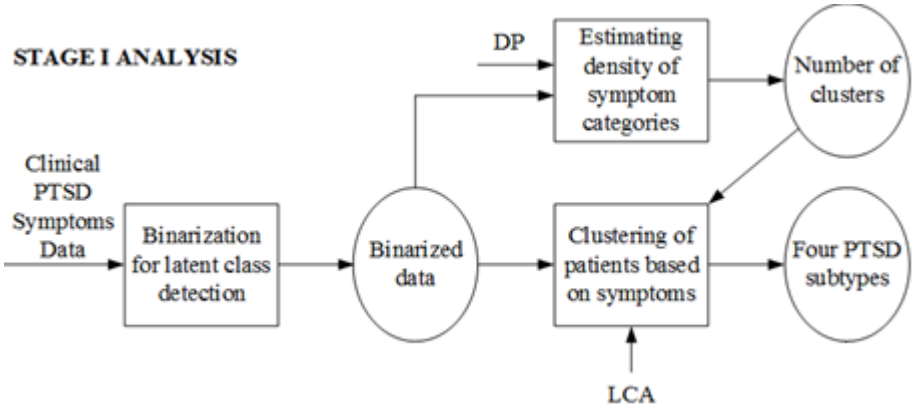


Figure 2. Process diagram of Bayesian-based framework. The proposed Bayesian framework is a detailed version of the Stage I analysis section of Figure 1. We begin stage I analysis by representing the data as binary data for analysis of symptom-absent (0) versus symptom-present (1) patients. A DP approach was used to first non-parametrically estimate the number of clusters appropriate for clustering the clinical data. We use these estimated number of clusters (four) into our LCA model to cluster our data into four PTSD subtypes. The proposed two-stage approach for future work is illustrated in **Appendix II**.

Clinical data

The clinical features of interest for our analysis focused on the 17 PSS scores corresponding to the 17 DSM-IV PTSD symptoms. The symptoms were measured based on a semi-structured interview questionnaire (see **Appendix I**) called Modified PTSD Symptom Scale (MPSS) that consist of 17 questions, corresponding to the 17 symptoms of PTSD, aimed at measuring the frequency and intensity of each symptom. The resulting PSS scores range from 0 to 3, denoting the absence (0), low presence (1), moderate presence (2) and extreme presence (3) of each symptom.

Analysis

Our analysis consists of a one-stage approach utilizing the clinical data. Stage I analysis (Bayesian-based cluster modeling) focuses on discovering subtypes by revealing hidden patient characteristics, based on the PTSD symptoms, using well-known clustering techniques.

Stage I Analysis: Bayesian-based Cluster Modeling

A latent class analysis (LCA) was used to discover PTSD subtypes in an unsupervised manner. The clinical data contains the 17 symptoms, re-labeled as symptom A-symptom Q, for all 78 subjects. These 17 polytomous outcome variables (i.e., symptoms with 4 PSS scores – 0,1,2,3) were dichotomized (i.e., symptoms with 2 PSS scores – 0,1) into two categories where scores equal to zero represent symptom-absent (0) patients and scores greater than zero represent symptom-present (1) patients. For this thesis, the variables were dichotomized in the interest of investigating the influence of brain connections on the absence/presence of a symptom, not in the interest of looking at the severity of PTSD (discussed in the *Further Research* section). These 17 dichotomous outcome variables were then used to cluster all 78 subjects. As a means to properly estimate the number of clusters to use in the LCA analysis, we incorporated a non-parametric Bayesian technique called the Dirichlet Process (DP).

Dirichlet Process

We implement a fit of the Dirichlet process model to perform density estimation on the dichotomous PSS symptom responses.[19] The statistical advantage of using the non-parametric prior from the Dirichlet model allows us to identify different distributions over the observed data and estimate underlying clusters of the data. This process helps overcome

the inherent heterogeneity present in PSS symptom responses. The function *DPdensity* in R was performed on the data to model the symptom scores using the default priors.[19] The resulting estimated number of clusters suggested the data should be clustered using four to five clusters. In this analysis, we performed our LCA analysis using four clusters based on the findings from a previous study but for future analyses we will conduct a comparative analysis using three to five clusters (as discussed in the *Further Research* section).[15]

Latent class analysis (LCA) model

LCA is a commonly used finite mixture model for subtype analysis of discrete and categorical variables. It assumes that there exists an underlying unobserved (or *latent*) categorical variable that stratifies observations into mutually exclusive latent groups or "classes".

Suppose $\boldsymbol{x}$ is the p-dimensional response pattern of the dichotomized PSS symptom responses for each patient. Denote the j^{th} symptom response of $\boldsymbol{x}$ as $\boldsymbol{x}_j$, where $j = 1,\ldots,p$ (p represents the total number of symptoms). Given a particular response pattern, the LCA defines the probability $P(\boldsymbol{X} = \boldsymbol{x})$ as follows:

$$P(\boldsymbol{X} = \boldsymbol{x}) = \sum_{k=1}^{K} \alpha_k \prod_{j=1}^{p} \prod_{r_j=1}^{R_k} \rho_{j,r_j|k}^{I(x_j = r_j)} \tag{1}$$

where K represents the total number of clusters (determined by DP), R_k represents the total number of responses for every symptom of the k^{th} cluster. The probability, ρ, is calculated using the EM algorithm.[20] It forms a matrix of dimension $p \times R_k$ for cluster k. Hence, for

cluster k, the $\rho_{j,r_j|k}^{I(x_j=r_j)}$ represents the probability that a patient with response $r_j \in \{0,1\}$ for the j^{th} symptom response in the clinical phenotype (e.g., anxiety, insomnia, or emotional numbness) $\boldsymbol{x}$. The coefficient α_k represents the latent subtype membership probability for each patient in cluster k. We also establish the indicator function $I(x_j = r_j)$ to be 0 or 1.[20]

The nature of the data (discrete, categorical) allows us to apply this latent class model on the PSS symptom responses to obtain PTSD subtypes. This is performed using the *poLCA* package in R, where the number of clusters is set equal to the number of clusters estimated in the DP analysis.[21]

Results

A one-staged analysis was performed to discover PTSD subtypes within the GTP cohort. We report results from our latent class cluster analysis where we group our patients based on the clinical data.

Clustering results

We implemented a latent class analysis on polytomous categorical PSS symptom responses, represented as dichotomous variables, for the sole purpose of clustering symptom-absent (0) and symptom-present (1) patients. A Dirichlet Process was implemented on the clinical data to estimate the number of clusters to specify in the LCA. The estimated number of clusters specified in the latent class model was four clusters. The analysis was then performed to cluster the subjects (n=78) based on the 17 dichotomized PSS symptom responses. We clustered the patients into four clusters with resulting groupings of 23 patients, 15 patients, 9 patients and 31 patients in each of the clusters, respectively.

Figure 3 illustrates the relationship between symptom responses (scores) and PTSD obtained the LCA model-based clustering. We were interested in identifying any significant symptom groupings based on the three main symptom categories of PTSD – intrusive, avoidance/numbness, and hyperarousal. The mean PSS scores were calculated for all symptoms across the subjects for each of the four clusters. Several interesting symptom-cluster characteristics were identified. The mean PSS scores closer to zero define those symptom-absent subjects; for example, in cluster 4, we can see the majority of the symptom-absent subjects, across all symptoms, were grouped together in this cluster. In particular, there were no subjects that displayed symptoms B, C and H as evident by the mean PSS scores of zero. Clusters 1-3 also contain significant cluster characterization. We can see that cluster 1 contains symptom-present subjects with moderate intrusive symptoms, low to high avoidance/numbness symptoms and low hyperarousal patients. Moreover, comparing both clusters 2 and 3, we observe these clusters contain subjects that display moderate to high symptoms in all three categories; especially, in cluster 3, where we see patients for certain symptoms – intrusive (A, C, D), avoidance/numbness (E, F, I, K), and hyperarousal (L, M, N) – who all display that particular symptom as indicated by a mean score of 1.

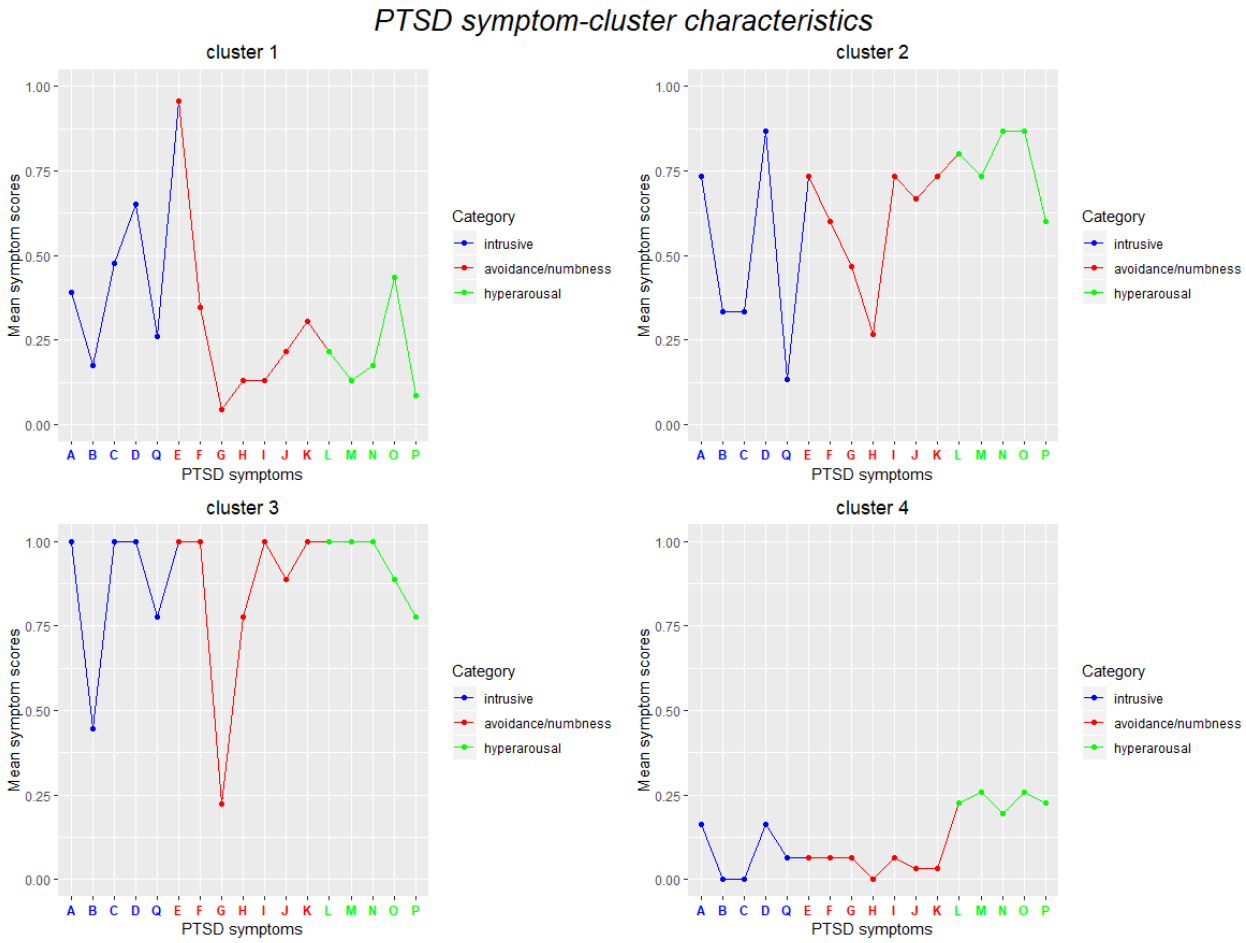


Figure 3. Mean symptom score characteristics for each cluster. The mean symptom scores for each cluster were plotted for the 17 PTSD symptoms. These 17 symptom scores can be stratified into three main groups: intrusive symptoms (A-D,Q), avoidance/numbness symptoms (E-K) and hyperarousal symptoms (L-P).

Visualizing the density and dispersion of clusters based on symptom categories is also valuable to cluster characterization. Figure 4 illustrates grouping of the four latent clusters based on the three main symptom categories – intrusive, avoidance/numbness, and hyperarousal – with the goal of

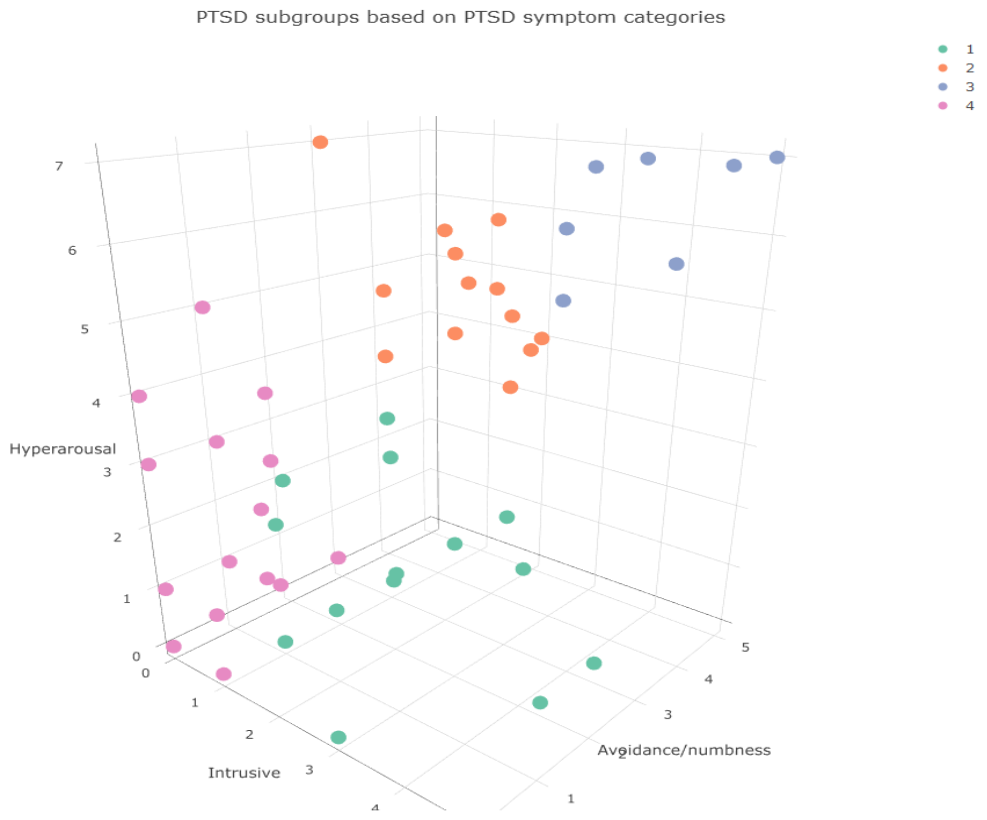


Figure 4. PTSD subgroups classified by three main symptom categories. The four resulting latent clusters of symptom-absent and symptom-present patients were classified based on the three symptom categories – intrusive, avoidance/numbness, hyperarousal. We plotted the latent-clustered patients (color-coded) against the symptom categories as represented by the x-axis, y-axis, and z-axis. This plot suggests clear distinctions (i.e., groupings) within the four clusters based on the three symptom categories.

investigating interesting symptom-category patterns within each of the four clusters. The classification results of the PTSD subtypes based on the three main symptom categories is shown in Figure 4. It is important to understand that points closer to the origin (0,0,0) define those symptom-absent patients and points further away define those symptom-present patients; hence, we see many cluster 4 patients near the origin and cluster 3 patients further away which aligns with Figure 3 where cluster 4 had mean symptoms scores close to 0 while cluster 3 had mean symptoms scores close to 1. This plot validates the performance of our clustering technique by highlighting the veracity behind symptom-category groupings within each cluster.

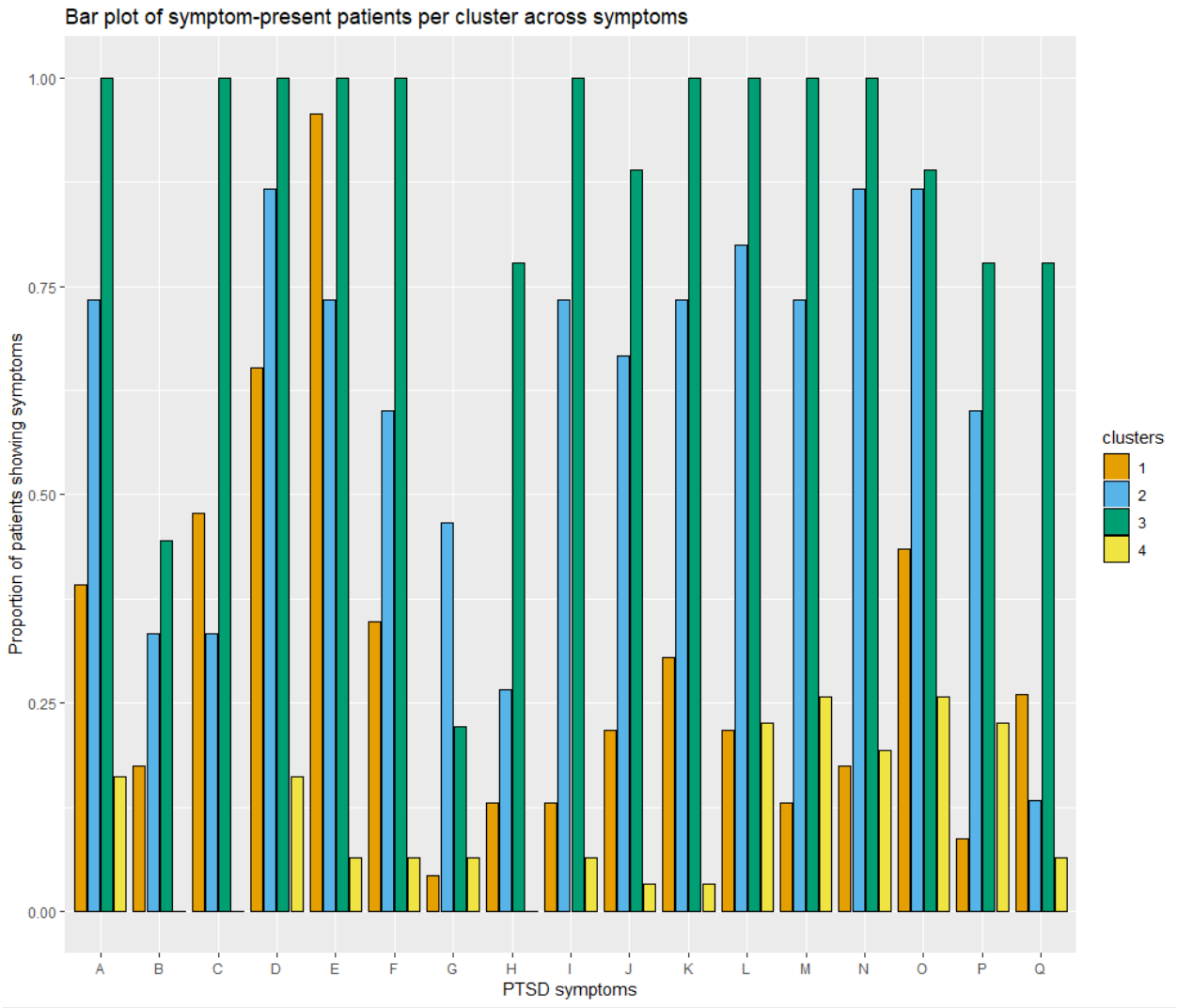


Figure 5. Proportion of symptom-present patients per cluster across symptoms. The proportion of patients who display symptoms were plotted for each cluster across the 17 PTSD symptoms. For example, about 75% of patients display symptom A in cluster 2. More importantly, this result suggests that symptom B has a smaller patient-proportion variance in patients across clusters, while symptoms C and E, for example, have larger patient-proportion variance in patients across clusters.

Figure 5 adds on to the characterization of our cluster groupings by allowing us to look at the patient-proportion variance for each symptom across clusters. Here, we generated patient proportion plots across symptoms for each cluster for the dichotomous outcome – symptom-absent (0) versus symptom-present (1) – patients. We can see that, in cluster 1,

the majority of the patients display symptom E (avoidance-related symptom which corresponds to question 5 of the MPSS found in **Appendix I**). Although identifying the most occurring symptom in a particular cluster is important, it is more interesting to identify the symptom-present patient proportion across symptoms in each cluster for identifying which symptoms vary the most across clusters and which symptoms vary the least across clusters. Thus, it is informative to investigate the difference in the height of the bars for each symptom as this provides us information of the patient-proportion variance; that is, symptom B seems to vary the least among clusters while symptoms E, L and N are seen to vary the most among clusters.

In summary, the characterization of cluster groups is useful for understanding the clinical significance behind the clusters. Here, we have performed group-wise characterization (i.e., *inter-cluster* characterization). However, it is important to consider the question of: what differentiates the patients *within* these clusters? That is, it is important to take into account an *intra-cluster* perspective where we use the brain networks of these patients and understand the neurobiological underpinnings that result in behavioral differences.

Discussion

This study examined cluster characteristics for better explaining the heterogeneity in PTSD symptom responses. A latent class analysis in conjunction with DP of PSS data drawn from a sample of socioeconomically disadvantaged, trauma-exposed females yielded evidence for distinct subtypes. We conduct subtype analysis to effectively explain the heterogeneity in our GTP cohort by using a one-staged systematic approach: use latent class analysis to learn PTSD subtype in an unsupervised manner.

PTSD subgroup identification

The latent class analysis used for clustering proved useful for revealing interesting cluster groupings using four clusters (determined by DP). Although heterogeneity exists among the PSS profiles (i.e. symptom responses) for each subject, the DP-based LCA discovered differences which were used to establish well-defined symptom subtypes based on the PSS profiles. Cluster 1 represents moderate to high symptom-present patients with avoidance-related symptom being predominantly displayed by the patients. Moreover, cluster 2 represents moderate- to high- symptom-present patients across all symptom categories. Additionally, cluster 3 represents very high symptom-present patients. Lastly, cluster 4 represents symptom-absent patients across the majority of the symptoms. This cluster performance facilitates the process of relating each subtype to the functional connectivity to further explain the inherent heterogeneity in PTSD symptoms.

Further research

The current thesis work illustrates a one-staged clustering approach. We aim to establish a two-stage approach (see **Appendix II**) where we incorporate network analysis to investigate underlying neurobiological causes for behavioral differences. We first discuss possible extensions related to Stage I and establish a proposal approach for Stage II.

Stage I Future Research

As discussed in the *Dirichlet Process* subsection of the Methods section, we performed our analysis using only four clusters, but there is room for further research in performing a comparative analysis using three clusters, four clusters and five clusters. We also performed our analysis representing the original polytomous outcome variables as

dichotomous variables for the purpose of analyzing two groups of patients: symptom-absent versus symptom-present patients. To build a more robust analysis, our next steps will be to repeat the analysis on the original polytomous outcome variables to see if we can obtain well-defined clusters as well as discover interesting brain network causes based on the severity of PTSD. Moreover, our current analysis involves three main symptom categories of PTSD – intrusive, avoidance/numbness, and hyperarousal – stratified into 17 symptoms. However, for future research, it will be important to incorporate the recent update in the DSM-V where avoidance and numbness are considered as two disparate categories; thus, analyzing on four categories instead of three categories of PTSD and working with 20 symptoms instead of 17 symptoms.

Stage II Future Research

As previously mentioned, to investigate the brain network causes for differential behavioral differences, we would like to incorporate a network analysis in conjunction with the cluster analysis. The following is a proposed extension to the one-stage approach: We will relate each subtype (i.e., the clusters identified by LCA with DP from Stage I) to the brain functional network for discovering connectivity patterns that will help us differentiate the different subgroups and contribute to overcome the inherent heterogeneity in our cohort study.

The brain network for each subject within a subtype will be estimated based on the brain regions from the PFC. There will be 12 regions (i.e. v = 12) and each region will have 146 time points (i.e. T = 146). If we define it as a matrix A_i (where A is the centered data with respect to the mean) of dimension 12 x 146 for subject *i*, then we can obtain a variance-covariance matrix $\sum_i$ of dimension 12 x 12 for subject *i* as follows: $\sum_i = A_i * A_i^T$, where the

operator (*) is matrix multiplication. We will then input this matrix to graphical least absolute shrinkage and selection operator (gLASSO) which will estimate a sparse inverse of the variance-covariance matrix $\Omega_i = \Sigma_i^{-1}$, based on a range of tuning parameters chosen empirically (discussed in the next section), where the observations (i.e., regions) for a given time point, *t*, follow: $y_t \sim N(0, \Omega_i)$. We will then apply a standard threshold of 0.005 to obtain the adjacency matrix and extract the upper-triangular matrix ($v(v\text{-}1)/2$) edge set for each subject within a subtype. This estimated network will be generated for a range of gLASSO tuning parameters ρ. Then for each of the 17 symptoms, we will select the tuning parameter that minimized the predictive mean squared error (MSE), as described next.

Selection of the tuning parameter ρ of gLASSO is a challenging task - a systematic way of selecting the tuning parameters will be proposed using the MSE obtained from a Bayesian binomial logistic regression analysis.

We will apply 10 empirically-evaluated ρ values evenly-split between the range of 0.001 to 0.04 to estimate brain networks. We will represent the estimated brain network, E_{ik}, as the vectorized upper-triangular adjacency matrix for every i^{th} subject within a subtype k where k=1,…,K. Likewise, we will represent the behavioral outcome, $y_{ijk\rho}$, for the j^{th} symptom of i^{th} subject in the k^{th} cluster for a particular ρ. We will then formulate a non-linear parametric relationship *f(x)* between the behavioral outcome and the brain network:

$$y_{ijk\rho} = f_{(\beta_{jk\rho})}(E_{ik\rho}) \quad (2)$$

where $\beta_{jk\rho}$ is the model parameter with dimensions 1 x $v(v\text{-}1)/2$ and its c^{th} element is represented by $\beta_{jk\rho c}$, where c=1,…,$v(v\text{-}1)/2$. There will be two requirements that must be

addressed in this model. The first requirement will be the selection of the mapping function *f*, where we consider the logistic link, and the second requirement will be the estimation of parameter $\beta_{jk\rho}$.

To estimate the parameter $\beta_{jk\rho}$, we will use the Bayesian binomial logistic regression with MCMC. The R package provides the *mlogit* function which inputs a dependent variable (the symptom response variable $y_{ijk\rho}$), independent variables (vectorized edge set variable $E_{ik\rho}$), and MCMC parameters, and then outputs the $\beta_{jk\rho}$ parameters, governed by the Bayesian binary logistic regression model.[22-24] This model assumes our latent variable, ω, follows a Polya-Gamma distribution $\omega \sim PG(b,0)$ with parameter $b > 0$, and when we integrate it out we get a resulting logistic regression model with a Gaussian prior on $\beta_{jk\rho}$ and we obtain our posterior samples for our $\beta_{jk\rho}$ using the Polya-Gamma method.[23] The number of MCMC iterations selected were M=5000, generating 5000 sets of $\beta_{jk\rho}$. We will then compute the average as follows $\frac{\sum_{j=1}^{M} \beta_{jk\rho}}{M}$ to obtain the average estimated beta parameters for the model that we presented in equation (2).

As previously mentioned, we will apply the logistic link; specifically, we will select the standard logistic regression for the mapping function *f* to generate the estimated symptom responses $\hat{y}_{ijk\rho}$ as probabilities based on the following model:

$$\hat{y}_{ijk\rho} = \frac{e^{\widehat{\beta}_{jk\rho} E_{ik\rho}}}{1 + e^{\widehat{\beta}_{jk\rho} E_{ik\rho}}} \tag{3}$$

Our goal now will be to calculate MSE values, according to $\frac{\sum_{k=1}^{K} \sum_{i=1}^{N_k} (y_{i,jk} - \hat{y}_{i,jk})^2}{N}$ for subtype k where N_k represents the total number of patients N in each subtype k. We will proceed to take the mean MSE across clusters to obtain our average MSE for each of our 10 ρ values per symptom. This will allow us to select the "optimal" tuning parameter for each symptom corresponding to the minimum average MSE. We will use the resulting "optimal" tuning parameters to estimate the brain network for the best predictive capability for a symptom.

For these estimated brain networks, we will calculate several graph-based metrics to characterize the PFC topological organization using the Brain Connectivity Toolbox in MATLAB.[24,25] We will consider (1) the density of the network, (2) the proportion of positive-valued and negative-valued edges, (3) the proportion of positive-valued and negative-valued estimated β's (4) the efficiency (global and local) of the network, (5) the characteristic path length (CPL) of the network, and (6) the mean clustering coefficient (MCC). The density and proportion of positive and negative edges provide inter-regional information, while the efficiency of a brain network measures how efficiently the nodes in the network communicate information. Global efficiency (GE) measures how quickly information of the entire network is exchanged between nodes and local efficiency (LE) measures the global efficiency of the nearest-neighbor nodes. The CPL holds information of the length of the path in a network, where a lower-valued number means shorter path length (i.e., path of highly-influencing network) and greater-valued number means longer path length. The GE and CPL are inversely-related since the longer the path length the lower the GE. Lastly, the MCC measures the average clustering of the brain regions (or nodes) in a network. In summary, these graph metrics will provide insightful brain network

information in relation to the symptoms since these features may highlight significant nodes predictive of displaying certain symptoms.

We will use the estimated brain networks to relate them to each symptom to discover network-influenced symptoms. Based on these patterns, we will generate predictive models that map the brain network to the symptoms. A heatmap, plots of the posterior means on the edge network, and estimated brain network plots will be used to visualize the influence of regional connectivity on the symptoms of PTSD. The investigation of connectivity patterns on PTSD symptoms will hopefully allow us to establish well-defined distinctions among the different symptoms that give rise to PTSD.

Limitations and Conclusions

It is important to consider the limitations present in this analysis. The clinical data analysis, exclusively, presents a few limitations. First, the demographics of the GTP data represents a very specific population - African American women – which gives rise to the issue of generalizability (i.e., this population study may not generalize to other populations). Second, trauma was assessed via retrospective self-report measures of PTSD symptoms so any conclusions of causality should be met with caution. Third, the scope of PTSD measures was limited to the DSM-IV 17-item MPSS, but recently, based on the DSM-V, the number of symptoms increased from 17 to 20 symptoms, thus it is important to consider these symptoms.[26]

The findings from this analysis should still be considered in light of these limitations. We have shown that clear distinctions lie within trauma-exposed patients. Moreover, we hope, with the incorporation of Stage II analysis, we will be able to show that network differences

exist for each PTSD symptom to better explain the underlying heterogeneity of the symptoms. The investigation of connectivity patterns on PTSD symptoms will hopefully allow us to establish well-defined distinctions among the different symptoms that give rise to PTSD; thus, facilitating the process for optimal diagnosis and treatment of PTSD.

References

1. Kessler, R. C., Sonnega, A., Bromet, E., Hughes, M., & Nelson, C. B. (1995). Posttraumatic stress disorder in the National Comorbidity Survey. *Archives of general psychiatry*, *52*(12), 1048-1060.
2. Breslau, N., Chilcoat, H. D., Kessler, R. C., & Davis, G. C. (1999). Previous exposure to trauma and PTSD effects of subsequent trauma: results from the Detroit Area Survey of Trauma. *American journal of Psychiatry*, *156*(6), 902-907.
3. Hoge, C. W., Castro, C. A., Messer, S. C., McGurk, D., Cotting, D. I., & Koffman, R. L. (2004). Combat duty in Iraq and Afghanistan, mental health problems, and barriers to care. *New England Journal of Medicine*, *351*(1), 13-22.
4. Stein, D. J., Seedat, S., Iversen, A., & Wessely, S. (2007). Post-traumatic stress disorder: medicine and politics. *The Lancet*, *369*(9556), 139-144.
5. O'donnell, M. L., Bryant, R. A., Creamer, M., & Carty, J. (2008). Mental health following traumatic injury: toward a health system model of early psychological intervention. *Clinical Psychology Review*, *28*(3), 387-406.
6. American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition*. Washington, DC: American Psychiatric Press; 1994.
7. Hidalgo, R. B., & Davidson, J. R. (2000). Posttraumatic stress disorder: epidemiology and health-related considerations. *The Journal of clinical psychiatry*.
8. Breslau, N. (2002). Epidemiologic studies of trauma, posttraumatic stress disorder, and other psychiatric disorders.
9. Heim, C., & Nemeroff, C. B. (2009). Neurobiology of posttraumatic stress disorder. *CNS spectr*, *14*(1 Suppl 1), 13-24.

10. VanElzakker, M. B., Dahlgren, M. K., Davis, F. C., Dubois, S., & Shin, L. M. (2014). From Pavlov to PTSD: the extinction of conditioned fear in rodents, humans, and anxiety disorders. *Neurobiology of learning and memory*, *113*, 3-18.
11. Shin, L. M., Wright, C. I., Cannistraro, P. A., Wedig, M. M., McMullin, K., Martis, B., ... & Orr, S. P. (2005). A functional magnetic resonance imaging study of amygdala and medial prefrontal cortex responses to overtly presented fearful faces in posttraumatic stress disorder. *Archives of general psychiatry*, *62*(3), 273-281.
12. Lukemire, J., Kundu, S., Pagnoni, G., & Guo, Y. (2017). Bayesian Joint Modeling of Multiple Brain Functional Networks. *arXiv preprint arXiv:1708.02123*.
13. Zhang, Q., Wu, Q., Zhu, H., He, L., Huang, H., Zhang, J., & Zhang, W. (2016). Multimodal MRI-based classification of trauma survivors with and without post-traumatic stress disorder. *Frontiers in neuroscience*, *10*, 292.
14. Gong, Q., Li, L., Tognin, S., Wu, Q., Pettersson-Yeo, W., Lui, S., ... & Mechelli, A. (2014). Using structural neuroanatomy to identify trauma survivors with and without post-traumatic stress disorder at the individual level. *Psychological medicine*, *44*(1), 195-203.
15. Rahman, A. F., Manatunga, A., Guo, Y., Peng, L., Warnock, M., Ressler, K. J., & Jovanovic, T. (2018). A latent class analysis of PTSD symptoms among inner city primary care patients. *Journal of psychiatric research*, *98*, 1-8.
16. Sun, D., Davis, S. L., Haswell, C. C., Swanson, C. A., Workgroup, M. A. M., LaBar, K. S., ... & Calhoun, P. S. (2018). Brain structural covariance network Topology in remitted Posttraumatic stress Disorder. *Frontiers in psychiatry*, *9*, 90.

17. Phillips, R. D., Wilson, S. M., Sun, D., Workgroup, V. M. A. M., & Morey, R. (2018). Posttraumatic stress disorder symptom network analysis in US military veterans: Examining the impact of combat exposure. *Frontiers in psychiatry*, *9*.
18. Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical science*, *23*(4), 439-464.
19. Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The annals of statistics*, 1152-1174.
20. Lanza, S. T., & Rhoades, B. L. (2013). Latent class analysis: an alternative perspective on subgroup analysis in prevention and treatment. *Prevention Science*, *14*(2), 157-168.
21. Linzer, D. A., & Lewis, J. B. (2011). poLCA: An R package for polytomous variable latent class analysis. *Journal of statistical software*, *42*(10), 1-29.
22. Polson, N. G., Scott, J. G., Windle, J., & Windle, M. J. (2016). Package 'BayesLogit'.
23. Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian inference for logistic models using Pólya–Gamma latent variables. *Journal of the American statistical Association*, *108*(504), 1339-1349.
24. Liu, J., Li, M., Pan, Y., Lan, W., Zheng, R., Wu, F. X., & Wang, J. (2017). Complex brain network analysis and its applications to brain disorders: a survey. *Complexity*, *2017*.
25. Complex network measures of brain connectivity: Uses and interpretations. Rubinov M, Sporns O (2010) NeuroImage 52:1059-69.

26. Pai, A., Suris, A., & North, C. (2017). Posttraumatic stress disorder in the DSM-5: Controversy, change, and conceptual considerations. *Behavioral Sciences, 7*(1), 7.

Appendix I: Modified PTSD Symptom Scale (MPSS)

SID - _____
Interviewer - _____

Modified PTSD Symptom Scale (MPSS)

The purpose of this scale is to measure the frequency and severity of symptoms in the past **TWO** weeks. Using the scale listed below, please indicate the frequency of symptoms to the left of each item.

FREQUENCY

0 Not at all **1** Once per week or less/ a little bit/once in awhile **2** Two to Four times per week/ somewhat/half the time **3** Five or more times per week very much/almost always

FREQUENCY

____1. Have you had recurrent or intrusive distressing thoughts or recollections about the event(s)?

____2. Have you been having recurrent bad dreams or nightmares about the event(s)?

____3. Have you had the experience of suddenly reliving the event(s), flashbacks of it, acting or feeling as it were re-occurring?

____4. Have you been intensely EMOTIONALLY upset when reminded of the event(s) (includes anniversary reactions)?

____5. Have you persistently been making efforts to avoid thoughts or feelings associated with the event(s) we've talked about?

____6. Have you persistently been making efforts to avoid activities, situations, or places that remind you of the event(s)?

____7. Are there any important aspects the event(s) that you still cannot recall?.

____8. Have you markedly lost interest in free time activities since the event(s)?

____9. Have you felt detached or cut off from others around you since the event(s)?

____10. Have you felt that your ability to experience emotions is less (e.g., unable to have loving feelings, do you feel numb, can't cry when sad, etc.)?

____11. Have you felt that any future plans or hopes have changed because of the event(s)? (e.g., no career, marriage, children, or long life?

____12. Have you been having persistent difficulty falling or staying asleep?

____13. Have you been continuously irritable or having outburst of anger?

____14. Have you been having persistent difficulty concentrating?

____15. Are you overly alert (e.g., check to see who is around you, etc) since the event(s)?

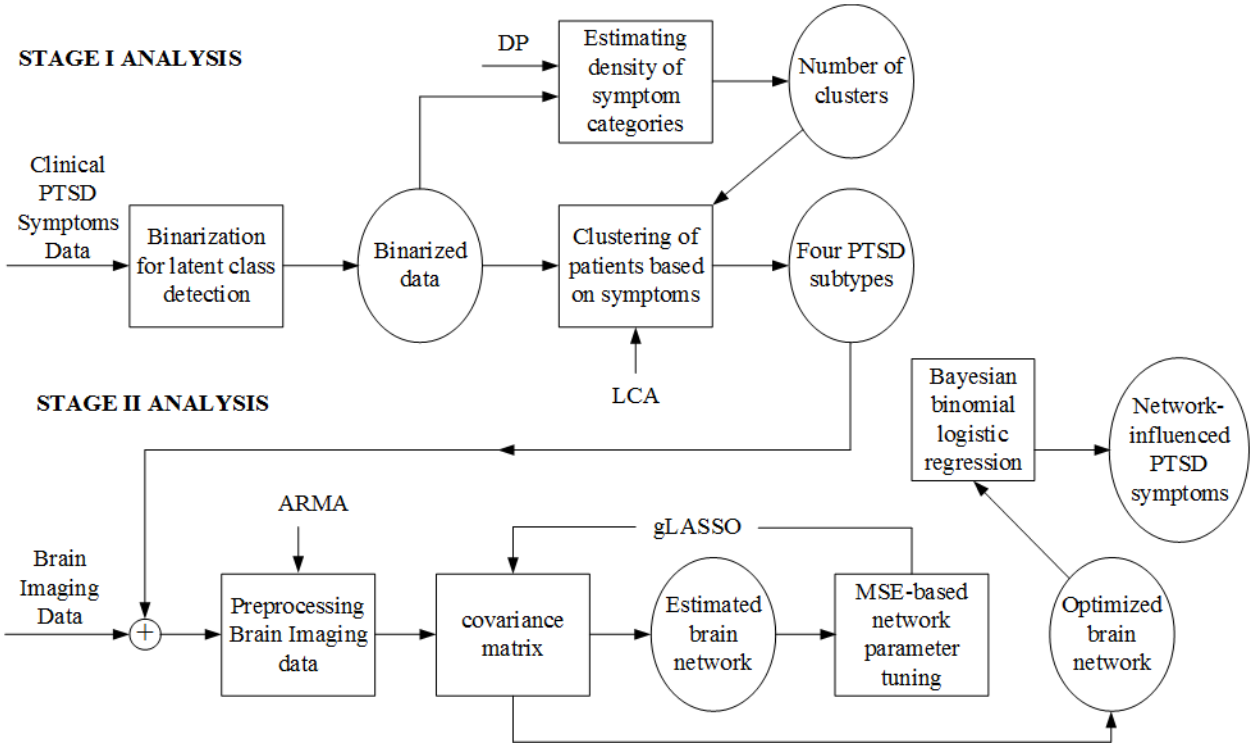
____16. Have you been jumpier, more easily startled, Since the event(s)?

____17. Have you been having intense PHYSICAL reactions (e.g., sweaty, heart palpitations) when reminded of the event(s)?

____18. How long have these symptoms bothered you?

Score **0**= < 1 month, **1**= 1-3 months, **2**= 3 months- 1 yr, **3**= > 1 yr

Appendix II: Proposed two-stage approach (Future Research)



Proposed Bayesian framework for investigating network-influenced PTSD symptoms. The proposed Bayesian framework is a detailed version of Figure 1. It involves a Stage I analysis and Stage II analysis using the clinical data and brain data, respectively. We begin stage I analysis by representing the data as binary data for analysis of symptom-absent (0) versus symptom-present (1) patients. A DP approach was used to first non-parametrically estimate the number of clusters appropriate for clustering the clinical data. We use these estimated number of clusters (four) into our LCA model to cluster our data into four PTSD subtypes. As discussed in *Stage II Further Research* We will begin stage II analysis by taking the respective brain imaging data of patients within each discovered subtype. We will first perform the standard preprocessing (e.g., autoregressive moving average (ARMA) model) of the brain imaging data for all subjects. We will then generate a covariance matrix of our regions of interest for all subjects, which we input into gLASSO with 10 empirically chosen ρ values. This outputs estimated brain networks for all subjects. We will then perform selection for the "optimal" tuning parameter using mean squared error (MSE) and proceed to obtain the optimal brain networks for every subject. We will then regress the symptom response on the optimal brain networks using a Bayesian binomial logistic regression to discover network-influenced PTSD symptoms.