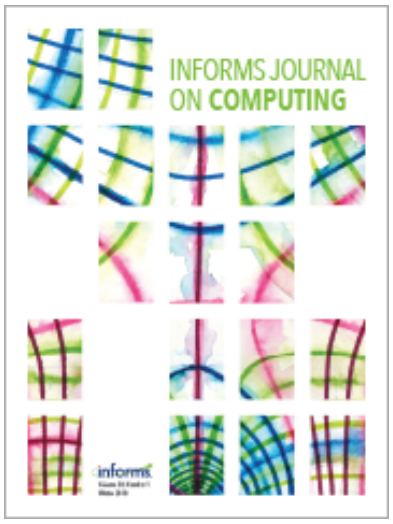


INFORMS Journal on Computing

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Bayesian Network Models for PTSD Screening in Veterans

Yi Tan, Prakash P. Shenoy, Ben Sherwood, Catherine Shenoy, Melinda Gaddy, Mary E. Oehlert

To cite this article:
Yi Tan, Prakash P. Shenoy, Ben Sherwood, Catherine Shenoy, Melinda Gaddy, Mary E. Oehlert (2023) Bayesian Network Models for PTSD Screening in Veterans. INFORMS Journal on Computing

Published online in Articles in Advance 28 Nov 2023

. https://doi.org/10.1287/ijoc.2021.0174

https://pubsonline.informs.org/journal/ijoc

INFORMS JOURNAL ON COMPUTING
Articles in Advance, pp. 1–15
ISSN 1091-9856 (print), ISSN 1526-5528 (online)

Bayesian Network Models for PTSD Screening in Veterans

Yi Tan,[a,*] Prakash P. Shenoy,[b] Ben Sherwood,[b] Catherine Shenoy,[b] Melinda Gaddy,[c] Mary E. Oehlert[c]

[a] College of Business, The University of Alabama in Huntsville, Huntsville, Alabama 35899; [b] School of Business, The University of Kansas, Lawrence, Kansas 66045; [c] VA Eastern Kansas Healthcare System, Leavenworth, Kansas 66048

*Corresponding author

Contact: yt0005@uah.edu, https://orcid.org/0000-0002-9235-0299 (YT); pshenoy@ku.edu (PPS); ben.sherwood@ku.edu (BS); catherine.shenoy@gmail.com (CS); Melinda.Gaddy@va.gov (MG); Mary.Oehlert@va.gov (MEO)

Received: July 8, 2021
Revised: March 13, 2022; October 18, 2022; April 26, 2023; August 18, 2023; August 26, 2023
Accepted: August 28, 2023
Published Online in Articles in Advance: November 28, 2023

https://doi.org/10.1287/ijoc.2021.0174

Abstract. The prediction of posttraumatic stress disorder (PTSD) has gained a lot of interest in clinical studies. Identifying patients with a high risk of PTSD can guide mental healthcare workers when making treatment decisions. The main goal of this paper is to propose several Bayesian network (BN) models to assess the probability that a veteran has PTSD when first visiting a U.S. Department of Veteran Affairs (VA) facility seeking medical care. The current practice is to use a five-question test called PC-PTSD-5. We aim to use the PC-PTSD-5 test, which is currently administered to most incoming new patients, and demographic information, military service history, and medical history. We construct a Bayes information criterion score-based BN, a group L_2-regularized BN (GL_2-regularized BN), and a naïve Bayes BN to assess the probability that a patient has PTSD. The GL_2-regularized BN is a new method for constructing a BN motivated by some of the challenges of analyzing this data set. A secondary goal is to identify which features are important in predicting PTSD. We discover that the following features help compute the probability of PTSD: PC-PTSD-5, service-connected flag, combat flag, agent orange flag, military sexual trauma flag, traumatic brain injury, and age.

History: Accepted by Ram Ramesh, Area Editor for Data Science & Machine Learning.

Supplemental Material: The software that supports the findings of this study is available within the paper and its Supplemental Information (https://pubsonline.informs.org/doi/suppl/10.1287/ijoc.2021.0174) as well as from the IJOC GitHub software repository (https://github.com/INFORMSJoC/2021.0174). The complete IJOC Software and Data Repository is available at https://informsjoc.github.io/.

Keywords: PTSD prediction • probabilistic classification • Bayesian network • regularization method • healthcare analytics

1. Introduction

Posttraumatic stress disorder (PTSD) is a prevalent and seriously impairing disorder, especially for veterans. Prediction of PTSD is a research domain that has attracted significant attention in the last two decades. Many studies focus on seeking the risk factors of PTSD (Brewin et al. 2000, Gaviria et al. 2016) that provide critical guidance for making individual mental healthcare decisions and help the U.S. Department of Veteran Affairs (VA) or other healthcare organizations to identify high-risk populations for PTSD. In recent years, machine learning techniques have been applied to PTSD prediction to fill in the gap between the scientific discovery of risk factors for PTSD and practical application in making accurate predictions of PTSD in individuals. Commonly used methods include support vector machine (SVM) (Galatzer-Levy et al. 2014), random forest (Schalinski et al. 2016), logistic regression (Holeva and Tarrier 2001), and naïve Bayes (Omurca and Ekinci 2015).[1]

VA facilities can use a PTSD prediction model as a screening tool to detect PTSD at an early stage. Given their experiences of combat and military sexual trauma, military veterans are at a higher risk for suffering from PTSD compared with their civilian counterparts (Norris and Slone 2013). Early detection of PTSD reduces the possible risk of untreated PTSD and makes the treatment more efficient. Traditional mental disorder screening relies mainly on psychological tests. The most commonly adopted screening instrument in VA hospitals since 2015 is the Primary Care PTSD Screen (PC-PTSD-5) from the fifth edition of the Diagnostic and Statistical Manual of Mental Disorders. Motivated by the need to predict PTSD for veterans more accurately at an early stage, we examine the construction of several Bayesian network (BN) models that incorporate the PC-PTSD-5 score and some other readily available information about a veteran, including demographic information, military service information, and the veteran's medical record.

BNs have been widely used in the medical domain because of their remarkable ability to handle uncertainty in the diagnostic process of diseases. Such models provide a natural way to handle missing data, allow data integration with domain knowledge, and facilitate learning about causal relationships between variables (Koller and Friedman 2009). Sun et al. (2011) presents a diagnostic system that assists doctors in diagnosing and tracking the development of mild cognitive impairment (MCI). Specifically, they deal with the problem of learning BN from incomplete data sets by introducing the mutual information and Newton interpolation Bayesian network (MNBN) algorithm. To characterize the functional relationships among symptoms of obsessive-compulsive disorder (OCD) and depression in patients with primary OCD, McNally et al. (2017a) estimate two networks (a regularized partial correlation network and a BN) using archival admission data from a hospital in Wisconsin. Using the same two approaches, McNally et al. (2017b) analyzes the PTSD symptoms in adults reporting childhood sexual abuse. Compared with the regularized partial correlation network, the BN can disclose potentially causal influence among symptoms. However, in practice, such influence between two symptoms can happen bidirectionally. One limitation of BN is that it disallows cycles whereby a symptom activates other symptoms that then loop back to influence the initial symptom. This paper proposes a screening tool to facilitate early detection and treatment of PTSD, where we do not have access to that advanced symptoms-related information. Instead, we consider using veterans' demographic information, military service, and medical records to see how this readily available information will affect the model's predictive power.

There are several methods for learning BN models from observational data. A BN model encodes conditional independence relations among the variables. One class of models, constraint-based methods, learns BNs by identifying conditional independence relations in a data set (Spirtes and Glymour 1991, Cheng et al. 2002). A feature of constraint-based methods is that they do not always result in a unique directed acyclic graph. More than one directed acyclic graph (DAG) may encode the same set of conditional independence relations in a data set. Typically these Markov equivalent DAGs are grouped and represented by a maximally partially acyclic graph (MPDAG), which contains both directed and undirected edges. As per Scutari et al. (2018), score-based algorithms (discussed next) are as accurate and computationally more efficient than constraint-based algorithms with large samples. Given a sample of more than 1 million veterans in our study, we did not construct a BN using constraint-based algorithms.

Another class of models, called score-based methods, learn BNs by optimizing scores such as Akaike information criterion (AIC), Bayes information criterion (BIC), and so on (Glover 1989, Tsamardinos et al. 2006). In this paper, first, we construct a BIC score-based BN. Next, we adopt a group L_2-regularized (GL_2-regularized) method to learn a BN model. Least absolute shrinkage and selection operator (LASSO) was first proposed by Tibshirani (1996) as a method that simultaneously performs feature selection and parameter estimation. We adapt a group version of LASSO to learn a BN model structure for predicting PTSD by using logistic regression to estimate the conditional probability distribution of each node. When fitting the model, we use known causal relationships to restrict directed arcs from causes to effects, which helps to improve our structure learning efficiency. For instance, we can rule out that PTSD causes a particular ethnicity. However, we do not interpret the relationships identified using the resulting model as causal discoveries. Besides, the group L_2 penalty is applied to guarantee the consistency of edge selection across multiple imputed data sets. Finally, we construct a naïve Bayes BN using only features identified as important from a feature importance study of the GL_2-regularized and the BIC score-based BNs.

We evaluate the prediction accuracies of all three BNs using a data set obtained from the Corporate Data Warehouse production domains of the VA Informatics and Computing Infrastructure (VINCI), which is a VA Health Services Research & Development (HSR&D) Resource Center that provides researchers with a nationwide view of detailed VA patient medical record data (U.S. Department of Veterans Affairs 2014). The data set consisting of patient-visit records of veterans has many missing values. A BN models the joint distribution for all variables in the data set. Fitting a BN model with missing data are as challenging as fitting a discriminative model, such as logistic regression, which directly models the conditional distribution of the response given the predictors. In our study, we deal with this issue by imputing missing values using Rubin's multiple imputation method (Rubin 1987). However, unlike discriminative models, a BN can easily make predictions even when we have missing predictors in the test set. Thus, making predictions from a BN will be more straightforward for the VA, where it is common for patients to have incomplete information. Out-of-sample testing shows that all three BN models are substantially better than using only the PC-PTSD-5 test.

Next, we explore the risk factors for PTSD among veterans. One advantage of using a BN model is that we have a transparent model of the linkages between features associated with PTSD. First, we identify the key features that directly influence the occurrence of PTSD in veterans using the Markov boundary of PTSD.[2] Second, we conduct a feature importance analysis to quantify how informative each feature is in predicting PTSD. The results from both studies suggest the same seven features important in predicting PTSD for military veterans.

1.1. Contributions of Our Study

The main contribution is a BN constructed by the GL_2-regularized method for predicting PTSD during the initial screening of veterans at a VA facility. The implementation code for our proposed method is available from the IJOC GitHub software repository (Tan et al. 2023). We also construct two other BNs—the BIC score-based BN and the naive Bayes BN—using existing methods to compare the accuracy of the GL_2-regularized BN. Second, our models can be used as a screening tool to detect PTSD at an early stage. Early detection of PTSD reduces its impact/risk and makes the treatment more efficient and less expensive. Our model substantially improves VA hospitals' prediction accuracy of the currently adopted PC-PTSD-5 test. Third, the BIC score-based and GL_2-regularized BNs identify important features in PTSD screening for military veterans. Identifying such important features of PTSD is helpful to the U.S. military in finding ways to avoid the problem and guide diagnosis, therapy, and disease control for VA hospitals. This also contributes to the literature by providing the first set of analyses on the modifying effect of patient demographic characteristics, military experience, and other comorbidities on the performance of screening tools in the primary care setting. Fourth, our paper introduces a new feature importance assessment method for BN models. We believe this method is cleaner than the commonly used permutation-based method.

The remainder of the paper is organized as follows. Section 2 discusses the primary motivation of this project—constructing a model to predict PTSD as a screening tool for patients. Section 3 describes the data set from VINCI and some of the challenges faced in constructing a BN. Section 4 describes a BIC score-based BN for PTSD screening. Section 5 describes the method we use to construct a GL_2-regularized BN. Section 6 presents the empirical evaluation of our proposed BN models and the corresponding feature importance analysis results. Finally, in Section 7, we summarize our findings and state some conclusions.

2. PTSD Screening Instrument

2.1. Background

PTSD occurs in people who have experienced or witnessed a traumatic event. It has been recognized as one of the most disabling psychopathological conditions affecting the U.S. veteran population. Veterans have a much higher prevalence of PTSD than nonveterans because of their increased exposure to life-threatening events. According to the National Center for PTSD, most PTSD in veterans varies by service era: between 11% and 20% for Operation Iraqi Freedom and Enduring Freedom, about 12% for the Gulf War, and about 15% for the Vietnam War. These numbers are significantly higher than U.S. civilians (about 7%–8%).

PTSD can disturb individual and family functioning, causing significant medical, financial, and social problems. Veterans with PTSD are more susceptible to sleep disorders, mood changes, reckless behavior, substance use disorder, and isolation. Traditional detection of veterans with PTSD relies to a large extent on primary care doctors.[3] Specifically, primary care doctors will refer to a mental health provider when they recognize symptoms of PTSD in a veteran. However, this procedure may delay the diagnosis and treatment due to the wait to see a primary care provider, followed by a wait to see a mental health provider. As per Penn et al. (2019), the average 2017 VA primary care wait time was 20 days. Also, doctors in primary care often prioritize physical health conditions and thus may overlook significant mental disorders, including PTSD, in the veterans they treat.

Early detection of PTSD reduces its risk, alleviates the related symptoms, and makes the treatment more efficient and less expensive. Numerous screening tests have been developed and adopted to assist clinicians in identifying the high risk for PTSD in VA hospitals. The screening was more commonly implemented first in primary care clinics. However, the VA has increasingly allowed veterans to go directly to mental health providers without being evaluated by a primary care doctor in recent years. A good screening test consisting of reasonable questions is easy for veterans to complete and has high sensitivity and specificity.

2.2. PC-PTSD-5

There are two steps in the PC-PTSD-5 screening test. First, subjects are asked if they have been exposed to a traumatic event. If the answer is no, the screening ends, and the subject receives a score of zero. If the answer is yes, they complete five additional yes/no questions about the traumatic event's impact on their life. The final score is the number of times the veteran responded yes to these five questions.

PTSD screening is currently required for *all* veterans receiving care in the VA. The VA started using PC-PTSD-5 in 2015. To facilitate early detection and treatment of PTSD veterans, the cutoff score in practice is determined as 3, which optimizes the test's sensitivity while maintaining the specificity at a minimum level of 0.8 (Prins et al. 2016). The PC-PTSD-5 is the most commonly used screening tool because it is brief and can be administered by a

primary care clinic staff; patients can easily understand and complete the questions. As a result, the PC-PTSD-5 exhibits very little sample selection bias.

One limitation of PC-PTSD-5 as a screening tool for PTSD is that it only focuses on the PTSD symptoms and disregards other factors, such as demographic characteristics and military experience. In this paper, using several BN models, we improve the screening power of PC-PTSD-5 with veterans' demographic information, historical military records, and traumatic brain injury (TBI) diagnosis. To the best of our knowledge, little work has been done to investigate the modifying effect of patient demographic characteristics, military experience, and other comorbidities on the performance of any screening tools in the primary care setting. Our study provides the first set of analyses on how much these factors contribute to improving the predictive power of PC-PTSD-5.

3. Data and Challenges

We use medical record data accessed through VINCI. We start by searching for veterans who have taken the PC-PTSD-5 test at least once by the end of 2019. Veterans' individual-level information is obtained by aggregating their lifetime visits to a VA facility. Our search identifies 1,113,676 distinct veterans.

3.1. Variable Definition and Miscellaneous Issues

Multiple descriptive epidemiological studies have been conducted to examine the patterns of PTSD in association with a range of demographic factors (Adams and Boscarino 2005, Gaviria et al. 2016). Following the literature, we start by collecting the veterans' *Age*, *Gender*, *MaritalStatus*, *Ethnicity*, *Religion*, and *Race*. Combat experience is a significant risk factor for the development of PTSD, whereas its impact varies depending on veterans' service era and if they have ever experienced any traumas. We also include military service variables, which can serve as indicators of veterans' combat status. Specifically, we include military exposures reported to the VA (*AgentOrangeFlag*, *IonizingRadiationFlag*, *SWAsiaConditionsFlag*); whether the veteran was treated for military service–related conditions including mental health disorders, musculoskeletal disorders, or other concerns (*ServiceConnectedFlag*); and the trauma types that they have experienced (*MilitarySexualTraumaFlag*, *CombatFlag*). These military experience–related variables will take the value of "yes" if a VA healthcare provider has selected this option to indicate that care was related to the relevant condition, exposure, or potential trauma. "No" is represented in the data set when a flag is relevant to the veteran (i.e., the veteran does have a condition, exposure, or experience that activates the flag in the medical record). However, no visits have ever been associated with treatment for that specified condition, exposure, or experience. There is a large number of empty values for these variables. For example, the proportion of empty values is 94.88% for *SWAsiaConditionsFlag* and 94.88% for *MilitarySexualTraumaFlag*. This is because these variables are irrelevant to most veterans. Thus, providers do not need to select either "Yes" or "No" associated with the veterans' care during their visits. For example, the *CombatFlag* is irrelevant and will always be empty for veterans who have never served in a combat theater or have not done so in the last several years. In this situation, the absence of a military service record is informative: A veteran is less likely to suffer from PTSD if the military experience information is missing than if the information has been collected. Accordingly, we treat the empty values as a separate value, called "unknown," for veterans' military service variables, instead of considering them as missing data.

Next, we collect the lifetime *PC-PTSD-5* test results. We use the highest scores for veterans who have taken the test multiple times. This is because PTSD is commonly a chronic disorder, and our goal is to predict whether a veteran has ever experienced PTSD in their lifetime. Finally, we investigate veterans' historical diagnosis of PTSD, our primary variable of interest, and traumatic brain injury (TBI), which has been widely recognized as a correlate of PTSD (Bryant 2011). These two variables will take the value "yes" if the veteran has been diagnosed with PTSD (or TBI respectively), and "no" otherwise.

Other data-collecting issues are discussed here.

- Inconsistent Records: For demographic information, we occasionally find inconsistent responses across different visits to the VA. In such situations, we treat the corresponding variables as missing.
- Numeric Variables: BN models work best with categorical variables. The most common way of addressing numeric variables in a BN is to assume that the conditional distribution of a numeric variable given its parents is normal (Lauritzen and Jensen 2001). However, this introduces bias when the normality assumption is violated. To handle numeric variables, such as *Age* and *PC-PTSD-5*, we discretize the variables using supervised discretization with a decision tree model (Kohavi and Sahami 1996). Specifically, we train a decision tree using *Age* (or *PC-PTSD-5*) to predict PTSD and use the splitting nodes in the fitted tree model as cutoff points for discretization. As a result, *Age* is discretized into three states: young (<53), middle (53–76), and old (>76), and *PC-PTSD-5* is discretized into

two states: Low (<3) and High (≥3). This result for *PC-PTSD-5* is consistent with the optimally sensitive cutoff score identified in previous research and used by the VA.

3.2. Summary Statistics

Table 1 provides the summary statistics of our variables. Given the sample of 1,113,676 veterans, 23.09% have been diagnosed with PTSD, and only 2.73% have been diagnosed with TBI. Most of the veterans are male (91.82%), white (71.38%), married (54.33%), not Hispanic or Latino (90.01%), and middle-aged (61.27%). It is worth noting that 23.39% of the data for religion is missing.

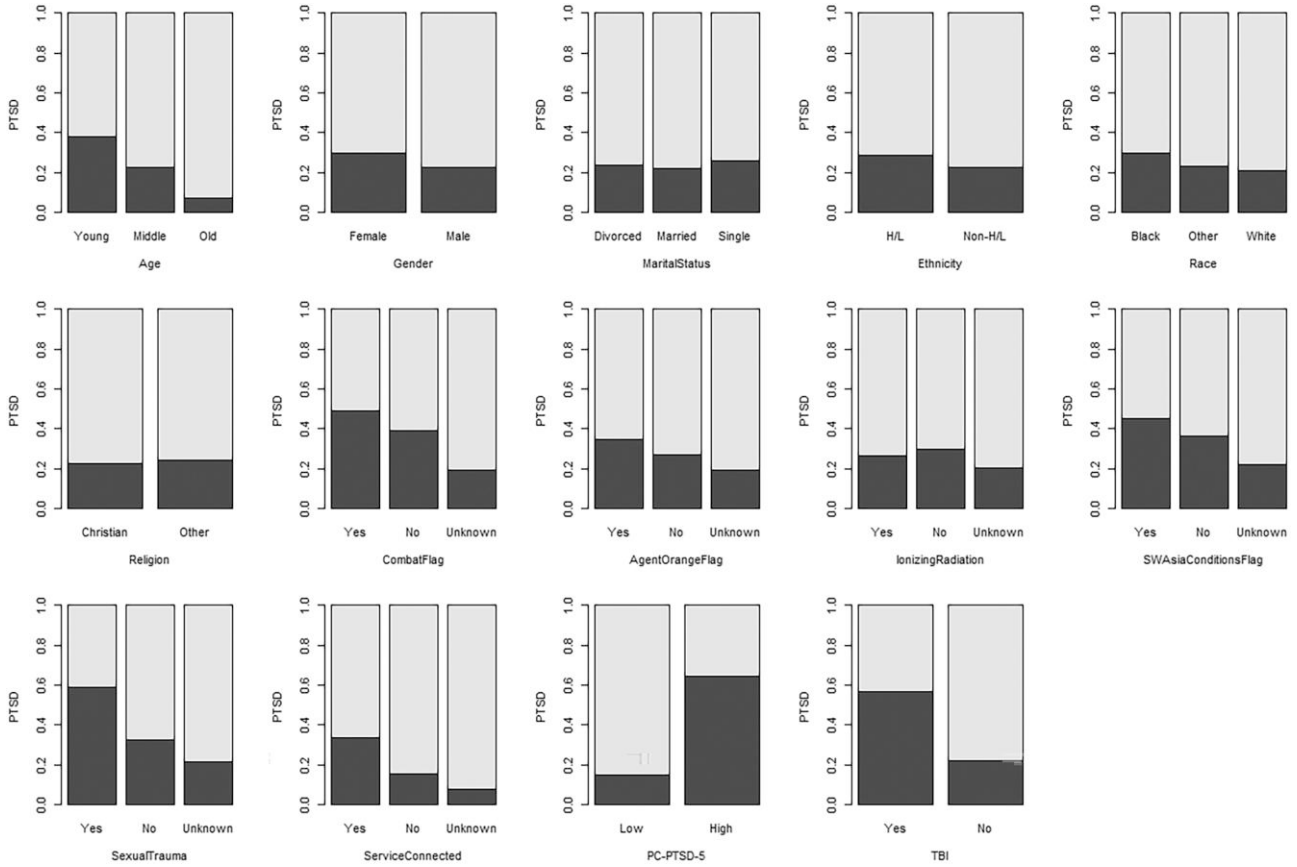
We show the prevalence of PTSD for our sample for different categories of other variables in Figure 1. The bar plots show that young, female, and Black veterans are more likely to suffer from PTSD. Also, being in combat, experiencing military sexual trauma, and service-connected trauma are associated with an increased risk of PTSD. Finally, a high score on the PC-PTSD-5 test is a strong indicator of PTSD.

Table 1. Summary Statistics for Key Variables

Features	States	Frequencies	Proportions (in %)
Age	Young	227,641	20.44
	Middle	682,393	61.27
	Old	203,642	18.29
Gender	Female	91,098	8.18
	Male	1,022,578	91.82
MaritalStatus	Divorced	350,143	31.44
	Married	605,050	54.33
	Single	152,772	13.72
	NA	5,711	0.51
Ethnicity	Hispanic/Latino	72,362	6.50
	Non-Hispanic/Latino	1,002,472	90.01
	NA	38,842	3.49
Race	Black	209,196	18.78
	Other	85,398	7.67
	White	794,939	71.38
	NA	24,143	2.17
Religion	Christian	801,801	72.00
	Other	51,382	4.61
	NA	260,493	23.39
CombatFlag	No	15,582	1.40
	Yes	126,059	11.32
	Unknown	972,035	87.28
AgentOrangeFlag	No	290,080	26.05
	Yes	120,519	10.82
	Unknown	703,077	63.13
IonizingRadiationFlag	No	308,323	27.69
	Yes	2,229	0.20
	Unknown	803,124	72.11
SWAsiaConditionsFlag	No	29,496	2.65
	Yes	27,552	2.47
	Unknown	1,056,628	94.88
MilitarySexualTraumaFlag	No	17,461	1.57
	Yes	39,517	3.55
	Unknown	1,056,698	94.88
ServiceConnectedFlag	No	32,027	2.88
	Yes	644,294	57.85
	Unknown	437,355	39.27
PC-PTSD-5	Low	932,380	83.72
	High	181,296	16.28
TBI	No	1,083,282	97.27
	Yes	30,394	2.73
PTSD	No	856,552	76.91
	Yes	257,124	23.09

Notes. We report the category values with corresponding frequencies and proportions. The category NA denotes missing data.

Figure 1. Prevalence of PTSD for Different States of Variables



Note. The height of a (white) black bar corresponds to the probability of (no) PTSD.

3.3. Some Challenges

We propose a supervised predictive model to bridge the gap between academic/clinical knowledge about PTSD and veterans' individual-level PTSD diagnosis. We identify four practical challenges and discuss how we plan to address them.

3.3.1. Challenge 1: Probabilistic Classification. The proposed model should be a probabilistic classifier that can predict the posterior probability of PTSD given states of some (or all) features. In this study, such probability represents a veteran's risk of suffering from PTSD. By changing the cutoff probability, our proposed method balances the tradeoff between sensitivity and specificity for different purposes, populations, or settings where a PTSD screening is conducted. For example, controlling the false-negative rate in primary care screening may be more important because it aims to facilitate early detection and treatment of PTSD cases that would otherwise go unrecognized. Support vector machines (SVMs) (Galatzer-Levy et al. 2014) are one popular machine learning technique used to predict the presence of PTSD. However, it is a deterministic approach that returns the class ($PTSD = Yes/No$) for each patient characteristic. The resulting model will have low adaptability to the varying PTSD screening purposes, populations, or settings. Other commonly used techniques include random forest (Schalinski et al. 2016), logistic regression (Holeva and Tarrier 2001), and naïve Bayes (Omurca and Ekinci 2015). In this paper, we explore BNs, which can predict the posterior probability of PTSD given observed values of features.

3.3.2. Challenge 2: Missing Data. A distinct advantage of a BN is that it can predict after only observing a subset of features. Specifically, some demographic information about veterans may be missing either because the information is not collected or not recorded in the medical record or the recorded data are inconsistent. For example, the information on *Religion* is missing for 23.39% of the veterans. Many other discriminative probabilistic classifiers, for example, (regularized) logistic regression, cannot predict from missing data without imputing them.

However, imputing the missing values biases the predicted probability, making it an unreliable measurement of the risk of PTSD. On the other hand, missing values in the training data are still an issue when learning a BN model, similar to learning discriminative models.

3.3.3. Challenge 3: Highly Skewed Features. In addition to missing data, highly skewed features are another issue that may reduce the sample's representativeness and bias the effect of PTSD screening, especially for minority groups.

This study addresses Challenges 2 and 3 by adopting a GL_2-regularized BN learning algorithm to reduce the model complexity. The multiple imputations (MI) method handles missing data in the training set for model learning.

3.3.4. Challenge 4: Large Search Space of Network Structures. Learning a BN structure from data are NP-hard because the network structure has to be a DAG. As our task is to construct a BN with 15 nodes from more than 1 million data instances, many commonly used methods are computationally intractable. In this study, we adopt an ordering-based search strategy and integrate it with domain knowledge to improve our structure learning algorithm efficiency.

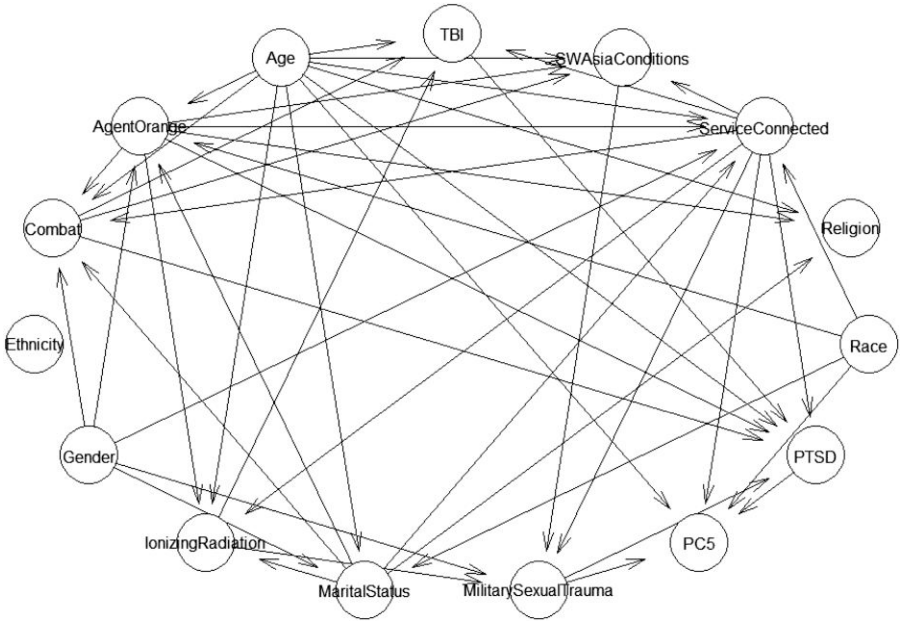
4. Bayes Information Criterion Score-Based BN

In this section, we investigate the construction of a BN model using the Bayes information criterion (BIC) score-based technique (Schwarz 1978).[4] The BIC score consists of a log-likelihood term that depends on the observed data and a penalty term that depends on the complexity of a BN model. Because each edge added to the BN increases the complexity of the model, this score-based term favors sparser network structures (with fewer edges for each node). Specifically, we minimize the BIC score using a hill-climbing greedy search and handle missing data with the structural expectation-maximization (EM) algorithm (Friedman 1998). Also, to prevent the strong influence of zero probabilities, parameters are estimated using the Laplace correction (Niblett 1987). We construct a BIC score-based BN using the *bnlearn* R package (Scutari 2010).

Figure 2 displays the resulting BIC score-based BN structure. As shown, the structure is relatively sparse with eight variables, *Age*, *AgentOrangeFlag*, *CombatFlag*, *MilitarySexualTraumaFlag*, *Race*, *ServiceConnectedFlag*, *TBI*, and *PC-PTSD-5*, in the Markov boundary of PTSD, and thus these variables provide the most direct predictive power.

Because our data set is quite large, typically we would expect to be able to fit BN models with even denser structures. However, in this study, because of missing data and highly skewed features, the observed training instances for estimating conditional probability tables are sometimes limited, making the corresponding estimation

Figure 2. Structure of the BN Model Constructed Using the BIC Score-Based Technique



unreliable. For example, to estimate the conditional probability table of PTSD, the number of veterans with *CombatFlag* = yes, *MilitarySexualTrauma* = yes, *TBI* = yes, and *Age* = old is zero. Also, the BIC score-based BN gives us no insight into the *relative* importance of features.

5. New Method for Constructing a GL_2-Regularized BN Model

Our proposed model is a regularized BN, built to address Challenges 1–4 described in Section 3. A BN is a directed acyclic graphical model with a set of m nodes $\{X_1, \ldots, X_m\}$. We define $Pa(X_j)$ as the vector of dummy variables corresponding to the variables that are the m_j parents of a node X_j. Given categorical data, we use a symmetric parametrization of the multinomial logistic regression for the conditional probability distribution of each node following Zhu and Hastie (2004):

$$P(X_j = x_{j,k} \mid Pa(X_j), \tilde{\boldsymbol{\beta}}_j) = \frac{\exp(Pa(X_j)'\tilde{\boldsymbol{\beta}}_{j,k})}{\sum_{l=1}^{K}\exp(Pa(X_j)'\tilde{\boldsymbol{\beta}}_{j,l})},$$

where $\tilde{\boldsymbol{\beta}}_j = (\tilde{\boldsymbol{\beta}}_{j,1}^T, \tilde{\boldsymbol{\beta}}_{j,2}^T, \ldots, \tilde{\boldsymbol{\beta}}_{j,K}^T)^T$ is the vector of unknown parameters to be estimated from data. Here K is the number categories for node X_j, and $\tilde{\boldsymbol{\beta}}_{j,k} = (\tilde{\boldsymbol{\beta}}_{j,k,0}, \tilde{\boldsymbol{\beta}}_{j,k,1}^T, \ldots, \tilde{\boldsymbol{\beta}}_{j,k,m_j}^T)^T$ corresponds to the kth category.

To tackle the challenge of structural learning of BN from high-dimensional data, Huang et al. (2012) proposed a sparse BN (SBN) structure learning algorithm. Given fully observed data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ where $\mathbf{x}_j$ is a $N \times df_j$ matrix of N observations and df_j degree of freedom for node X_j, we define X_{-j} as the set of all nodes excluding X_j. Assuming all these nodes are candidate parents of node X_j, we have $m_j = m - 1$. The model parameters can be reorganized as $\boldsymbol{\beta}_j = (\boldsymbol{\beta}_{j,0}^T, \boldsymbol{\beta}_{j,1}^T, \ldots, \boldsymbol{\beta}_{j,j-1}^T, \boldsymbol{\beta}_{j,j+1}^T, \ldots, \boldsymbol{\beta}_{j,m}^T)^T$, where $\boldsymbol{\beta}_{j,0} \in \mathbb{R}^{df_{j0}}$ is the vector of intercepts and $\boldsymbol{\beta}_{j,i} \in \mathbb{R}^{df_{ji}}$ is the vector of regression coefficients of parent node X_i. Then the problem is formulated as obtaining a sparse estimate of $\boldsymbol{\beta}_j$'s under the constraint that the estimated BN structure G must be a DAG. The nonzero element of $\boldsymbol{\beta}_j$s indicates the presence of edges in the structure G. Specifically, the estimate of $\boldsymbol{\beta}_j$'s is obtained by minimizing the negative log-likelihood (NLL) of each node, with the sparsity enforcing l_1 penalty as

$$\min_{\boldsymbol{\beta}_j} \frac{1}{N}\sum_{j=1}^{m} NLL(\mathbf{x}_j, \mathbf{x}_{-j}, \boldsymbol{\beta}_j) + \lambda \sum_{j=1}^{m}\sum_{i\neq j} \|\boldsymbol{\beta}_{j,i}\|_1 \quad s.t. \quad G \in DAG. \tag{1}$$

Given the estimate of $\boldsymbol{\beta}_j$s, the set of parents for node X_j can be found as $Pa(X_j) = \{X_i \mid \|\boldsymbol{\beta}_{j,i}\|_1 \neq 0\}$. Tuning parameter λ denotes the strength of regularization and can be determined by out-of-sample prediction performance in a validation set.

5.1. Multiple Imputation

This section describes MIs used to address the training set's missing data. MI was first proposed by Rubin (1987) and has been widely used in large-scale healthcare/medical studies (Rubin and Schenker 1991, Van Buuren 2007). It has the practical advantages of preserving sample size and statistical power and allowing standard complete-data methods of analysis to be used. Multiple imputations provide unbiased parameter estimates if the missing structure is missing completely at random or missing at random; whether an observation is missing or not does not depend on the value of the missing data. MI has three basic steps: (1) create D different data sets by imputation; (2) analyze each of the D completed data sets; and (3) integrate the D analysis results into a final result.

We conduct $D = 4$ imputations in this study.[5] Specifically, we compute the posterior probability for each node with missing values using BIC score-based BN as shown in Figure 2. Then the missing entries are imputed by randomly generating new data from the corresponding posterior probability distribution.

When multilevel features and responses are present, the LASSO penalty may not be satisfactory as it only selects individual variables instead of whole factors. The LASSO solution also depends on how the dummies are encoded. Yuan and Lin (2006) propose group LASSO to overcome these issues. Chen and Wang (2013) extend this idea to multiple-imputed data to ensure consistency in variable selection across different imputed datasets. In this paper, we denote $\hat{\boldsymbol{\beta}}_j^1, \ldots, \hat{\boldsymbol{\beta}}_j^D$ as the vectors of estimated regression parameters for child node X_j on the D imputed data sets, where $\hat{\boldsymbol{\beta}}_j^d = (\hat{\boldsymbol{\beta}}_{j,0}^{d^T}, \hat{\boldsymbol{\beta}}_{j,1}^{d^T}, \ldots, \hat{\boldsymbol{\beta}}_{j,j-1}^{d^T}, \hat{\boldsymbol{\beta}}_{j,j+1}^{d^T}, \ldots, \hat{\boldsymbol{\beta}}_{j,m}^{d^T})^T$. Then, the model parameters are extended as $\boldsymbol{\beta}(D)_{j,i} = (\boldsymbol{\beta}_{j,i}^{1^T}, \ldots, \boldsymbol{\beta}_{j,i}^{D^T})^T \in \mathbb{R}^{df_{ji}\cdot D}$. If X_i is important for predicting X_j, $\hat{\boldsymbol{\beta}}_{j,i}^d$s should be all nonzero, and if X_i is not important for predicting X_j, $\hat{\boldsymbol{\beta}}_{j,i}^d$s should be all zero for any given imputed data set d. Thus, we estimate the model by minimizing

$$\min_{\boldsymbol{\beta}_{j,i}^d} \frac{1}{N\cdot D}\sum_{j=1}^{m}\sum_{d=1}^{D} NLL(\mathbf{x}_j^d, \mathbf{x}_{-j}^d, \boldsymbol{\beta}_{j,i}^d) + \lambda \sum_{j=1}^{m}\sum_{i\neq j} \sqrt{p_{j,i}}\|\boldsymbol{\beta}(D)_{j,i}\|_2 \quad s.t. \quad G \in DAG. \tag{2}$$

The L_2 norm $\|\boldsymbol{\beta}(D)_{j,i}\|_2$ is called the group LASSO penalty, and $p_{j,i} = df_{ji} \cdot D$ is the varying group size. The penalty function is adjusted by $\sqrt{p_{j,i}}$ to apply the same penalization to large and small groups. The group LASSO penalty guarantees the consistency of edge selection concerning all different predictor levels, response levels, and imputed data sets. The consistency in edge selection ensures the same network structure across different imputed data sets. The final estimation of $\hat{\boldsymbol{\beta}}_{j,i}$ is calculated based on Rubin's rule as $\hat{\boldsymbol{\beta}}_{j,i} = 1/D * \sum_{d=1}^{D} \hat{\boldsymbol{\beta}}_{j,i}^{d}$.

5.2. Ordering-Based Search

Solving the optimization in Equation (2) is a challenge, given the constraint that the estimated BN structure G must be a DAG because of the vast search space of network structures. Much work has been done to address this problem, but only a few outperform the baseline of greedy hill-climbing with tabu lists.[6] In this paper, we adopt an ordering-based search strategy (Teyssier and Koller 2005) and use the greedy hill-climbing search with a tabu list. Determining an appropriate ordering is a complex problem. However, our causality-based clinical knowledge helps significantly reduce our search space.

We conduct an ordering-based search by seeking the best ordering $\prec$ over $X_1, \ldots, X_m$, such that if X_i is a potential candidate for $Pa(X_j)$, then $X_i \prec X_j$. Once the ordering $\prec$ is determined, finding the optimal BN consistent with $\prec$ is no longer NP-hard because we can easily implement group LASSO on each node separately. We use hill-climbing to find $\prec$, that is, only consider swapping a pair of adjacent nodes in the ordering for each move until the value of Objective Function (2) does not decrease:

$$(\ldots, X_{i-1}, X_i, X_{i+1}, X_{i+2}, \ldots) \rightarrow (\ldots, X_{i-1}, X_{i+1}, X_i, X_{i+2}, \ldots).$$

There are only two new neighborhoods generated, (X_{i-1}, X_{i+1}) and (X_i, X_{i+2}), for each move. We use a tabu list to prevent the algorithm from reversing a swap executed recently in the search.

We use domain knowledge to reduce the search space of possible ordering $\prec$. Specifically, based on causal domain knowledge, we divide the nodes into five layers, illustrated in Figure (3). For example, individual characteristics at the first layer are attributes the other nodes cannot cause. As TBI is usually present before PTSD, we constrain TBI to precede PTSD. If a node X_i is at the preceding layer of node X_j, then X_i should always precede X_j in $\prec$. Given the ordering, define $\mathbf{x}_{i \prec j}$ as the set of potential parents for node j. In this way, we restrict our ordering search space only within the second layer.

With a predetermined ordering $\prec$, Equation (2) can be transformed as

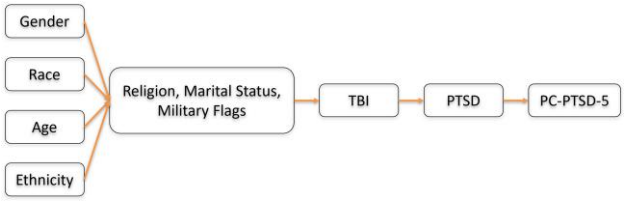
$$\min_{\boldsymbol{\beta}_{j,i}^{d}} \sum_{j=1}^{m} \left[\frac{1}{N \cdot D} \sum_{d=1}^{D} NLL(\mathbf{x}_j^d, \mathbf{x}_{i \prec j}^d, \boldsymbol{\beta}_{j,i}^d) + \lambda \sum_{i \prec j} \sqrt{p_{j,i}} \|\boldsymbol{\beta}(D)_{j,i}\|_2 \right]. \tag{3}$$

Then our task becomes solving m optimization problems independently. This allows us to easily apply the sparse group LASSO algorithm proposed by Vincent and Hansen (2014). Online Appendix D summarizes our model training procedure with a given value of regularization parameter λ.

6. Empirical Results of BN Models

This section presents the proposed BN models' prediction performance and feature importance. All the experiments were performed on the VINCI Workspace, using R. Because the original data are confidential, we provide a synthetic data set along with the R scripts that were used for all numerical experiments presented in this section on the IJOC GitHub software repository (Tan et al. 2023).

Figure 3. (Color online) Ordering-Based Search



Notes. Arrows represent constraints on directed edges. There can be no edges from a variable in the next level to a variable in the preceding level. All variables preceding PTSD can be potential parents of PTSD.

The overall performance is evaluated in terms of mean squared error (*MSE*), defined as follows:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\{\hat{P}(PTSD_i = \text{“Yes”}) - \mathbf{1}(PTSD_i = \text{“Yes”})\}^2.$$

Here N is the testing sample size, $PTSD_i$ is the observed PTSD status for the ith testing observation, $\hat{P}(PTSD_i =$ “Yes”) is the predicted probability that the ith testing observation is suffering from PTSD, and $\mathbf{1}()$ is an indicator function for the condition in the parentheses.

In practice, the use of screening tools is dependent on their cutoff point. An optimal cutoff point will provide a good balance of sensitivity–specificity tradeoffs concerning different screening purposes, populations, or settings. Accordingly, we also investigate false-negative (*FN*) and false-positive (*FP*) rates as follows:

$$FN = 1 - \frac{\sum_{i=1}^{N}\mathbf{1}(PTSD_i = \text{“Yes”})\cdot\mathbf{1}(\hat{PTSD}_i = \text{“Yes”})}{\sum_{i=1}^{N}\mathbf{1}(PTSD_i = \text{“Yes”})},$$

and

$$FP = 1 - \frac{\sum_{i=1}^{N}\mathbf{1}(PTSD_i = \text{“No”})\cdot\mathbf{1}(\hat{PTSD}_i = \text{“No”})}{\sum_{i=1}^{N}\mathbf{1}(PTSD_i = \text{“No”})},$$

where $\hat{PTSD}_i$ is the predicted PTSD status for testing observation i.

We randomly divide our data set into three parts, a training set with 913,676 instances, a validation set with 100,000 instances, and a test set with the remaining 100,000 instances. In the training set, the method described in Section 5 (with a prespecified sequence of tuning parameter values) is used to train the model. For variables with missing values, we impute them four times using the BIC score-based BN model. As our goal is to measure veterans' risk of suffering PTSD, we compare the *MSE* of models trained with different values of λ in the validation set and determine the optimal value of λ based on it. Finally, we assess the predictive accuracy of the resulting model using data in the test set.

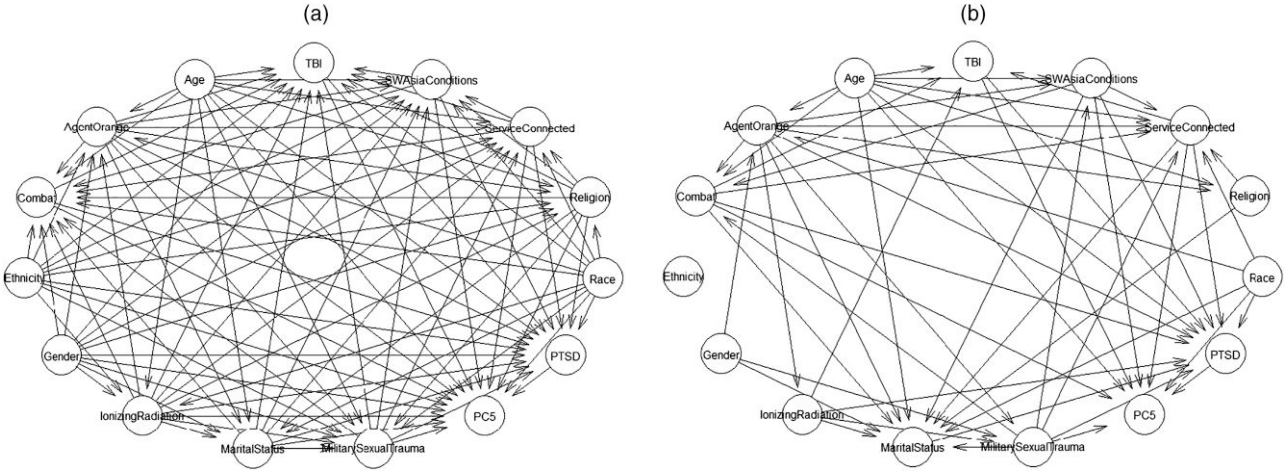
Table 2 summarizes the variables in the Markov boundary of *PTSD* given models generated with different λ values. As the value of λ increases, *Ethnicity*, *Religion*, and *Gender* are the nodes that leave the Markov boundary of PTSD early and tend to have the least significant predictive power. Conversely, *PC-PTSD-5*, *ServiceConnectedFlag*, and *CombatFlag* are still in the Markov boundary with $\lambda = 0.03$ and therefore contribute the most in predicting PTSD.

To optimize the prediction accuracy, we pick $\lambda_{min} = 0.00005$, which is associated with the smallest MSE in the validation set.[7] The resulting BN structure is shown in Figure 4(a). This model exhibits a very dense structure, probably due to our large training data set. As shown in Table 2, all 14 variables are in the Markov boundary of *PTSD*, indicating that they are informative in predicting PTSD.

Table 2. Summary of the Variables in the Markov Boundary of *PTSD* Given Models with Different Values of the Turning Parameter

	λ											
	0.00001	0.00003	0.00005	0.0001	0.0003	0.0005	0.001	0.002	0.003	0.005	0.01	0.03
PC-PTSD-5	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
ServiceConnectedFlag	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CombatFlag	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Age	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
TBI	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
AgentOrangeFlag	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
MilitarySexualTraumaFlag	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	
SWAsiaConditionsFlag	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Race	✓	✓	✓	✓	✓	✓	✓	✓	✓			
MaritalStatus	✓	✓	✓	✓	✓	✓	✓	✓	✓			
IonizingRadiationFlag	✓	✓	✓	✓	✓	✓	✓	✓	✓			
Gender	✓	✓	✓	✓	✓	✓	✓					
Religion	✓	✓	✓	✓	✓	✓	✓					
Ethnicity	✓	✓	✓	✓	✓	✓						

Figure 4. Structure of the GL_2-Regularized BN Model at (a) $\lambda_{min} = 0.00005$ and (b) $\lambda_{1se} = 0.002$



A comparison of our proposed GL_2-regularized BN with BIC score-based BN in terms of *MSE* is reported in Table 3. The results show that the GL_2-regularized BN with λ_{min} performs slightly better than the BIC score-based BN, although the difference may not be significant.

6.1. Cutoff Value of Probability of PTSD

The practical use of a PTSD screening tool is highly dependent on its cutoff value. Therefore, we investigate the optimal cutoff value for PTSD prediction by striking a proper balance between sensitivity and specificity. A commonly used technique for deciding the cutoff point in primary care screening is to maximize the sensitivity score while maintaining the specificity at a prespecified level, ensuring that the screening tool will not capture too many false positives. We first set the minimum specificity in this study to be 0.80 (Prins et al. 2016). As a consequence, the optimal cutoff for our proposed GL_2-regularized BN with λ_{min} is identified to be 0.2616 with the corresponding sensitivity as 0.7039, specificity as 0.8000.[8] Compared with BIC score-based BN, the sensitivity of GL_2-regularized BN with λ_{min} is 1.08% higher. We also investigate the sensitivity and specificity using only PC-PTSD-5 with a cutoff of three. The PC-PTSD-5 test provides a sensitivity of 0.4599 and a specificity of 0.9243. Suppose VA wants to maintain the same level of specificity. In that case, the optimal cutoff for our proposed GL_2-regularized BN is 0.4154, which results in a sensitivity of 0.4971 and a specificity of 0.9247. This indicates a 3.72% increase in screening sensitivity compared with the PC-PTSD-5 test. As the number of new VA enrollees is at least around 160,000 each year (FY 2017), this translates to 1374 more successfully detected PTSD veterans per year.[9] Also, GL_2-regularized BN's sensitivity is 0.13% higher than that for BIC score-based BN, which translates to almost 50 more successfully detected PTSD veterans per year.[10] Accordingly, our proposed GL_2-regularized BN consistently performs better in these metrics. A 0.13% (almost 50 veterans per year) difference does not appear significant, but it is a noticeable improvement over a long period.

6.2. Feature Importance

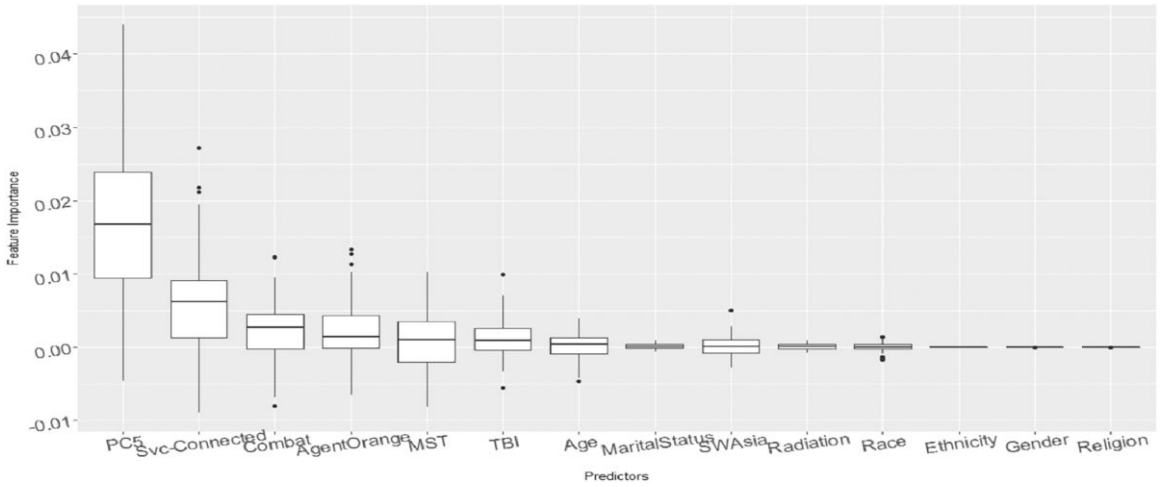
One disadvantage of the model with λ_{min} is that it lacks interpretability given such a complicated network structure. To provide a better interpretation of how each of these variables affects the risk of PTSD, we also report the BN model with $\lambda_{1se} = 0.002$, which is the largest λ value with its *MSE* = 0.1271 within one standard error of λ_{min}. The main idea of this "one-standard-error" rule is to choose the simplest model whose accuracy is comparable with the best one. The model structure for λ_{1se} is shown in Figure 4(b), which is much sparser than that for λ_{min}.

Table 3. Summary of Results

	$BN_{GL_2-\lambda_{min}}$	$BN_{GL_2-\lambda_{1se}}$	BN_{BIC}	NB
MSE	0.1263	0.1268	0.1265	0.1337
$FN_{spec-0.8}$	0.2961 (0.2000)	0.2994 (0.1988)	0.3069 (0.1922)	0.3121 (0.1911)
$FN_{spec-pc5}$	0.5029 (0.0753)	0.5041 (0.0757)	0.5042 (0.0756)	0.5381 (0.0664)

Note. The corresponding *FP* for $FN_{spec-0.8}$ and $FN_{spec-pc5}$ are in parentheses.

Figure 5. Feature Importance Box Plots Based on the GL_2-Regularized BN Model with $\lambda = \lambda_{1se}$



Note. The features are sorted by decreasing values of the median.

Eleven variables, *Age*, *AgentOrange*, *Combat*, *IonizingRadiation*, *MaritalStatus*, *MilitarySexualTrauma*, *Race*, *ServiceConnected*, *SWAsiaConditions*, *TBI*, and *PC5*, are still in the Markov boundary of PTSD. Besides, as shown in Table 3, the prediction performance of this model is comparable with GL_2-regularized BN model with $\lambda = \lambda_{min}$, and BIC score-based BN.

To better understand the usefulness of features in PTSD screening, we check the importance of each feature based on the GL_2-regularized BN model with $\lambda = \lambda_{1se}$. Precisely, we measure the importance of a feature as the increase in *MSE* when the feature values are removed from the test set. This provides a numerical measure of how much each variable contributes to predicting PTSD.

Typically, given a model, the importance of a feature is assessed by a random permutation of the feature values in the test set (Altmann et al. 2010). This is done because most machine learning methods cannot be used with missing feature values. Because BN models can, we propose a new approach to calculate a feature's importance by removing the feature column from the test set. This is cleaner than permuting the values of the feature, which may introduce noise in the prediction process. Earlier in the paper, we used multiple imputations to train a BN model with missing variables. In this section, we consider a missing variable for prediction purposes to test a feature's importance. Although BN models can make predictions with missing values, prediction performance will suffer if missing an important feature.

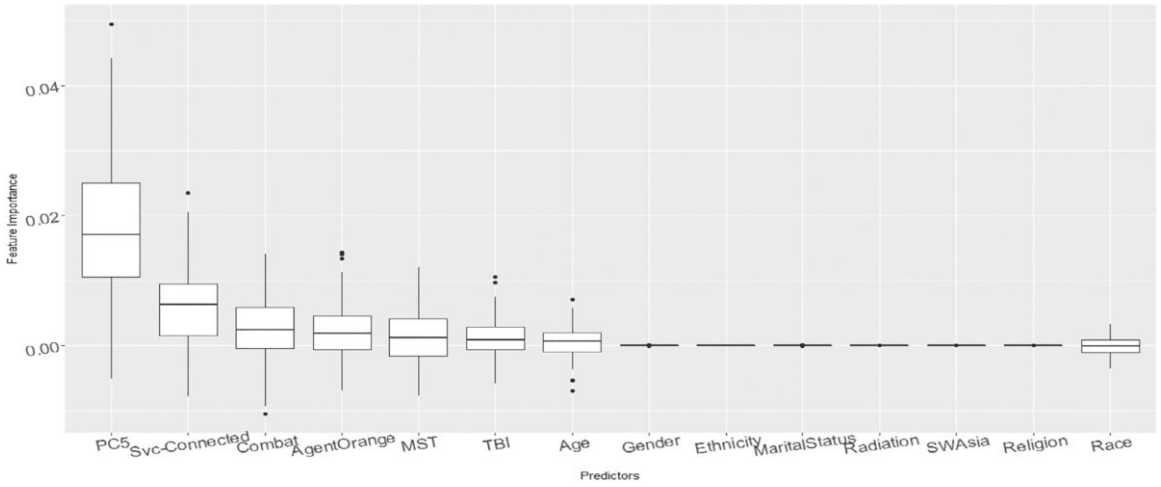
Making exact inferences with BN is computationally inefficient, especially given such a large data set. Instead of evaluating the increase of *MSE* using the entire test set, we analyze it with random sampling. Specifically, we randomly sample 100 veterans from the test set with replacement, calculate the increase of *MSE* for each predictor assuming its value is missing, and repeat these two steps 500 times. The feature importance for the proposed GL_2-regularized BN model with λ_{1se} is shown in Figure 5. Unsurprisingly, *PC-PTSD-5* has the highest feature importance among all 14 features. Also, *ServiceConnectedFlag*, *CombatFlag*, *AgentOrangeFlag*, *MilitarySexualTraumaFlag*, *TBI*, and *Age* are all important features that help to predict PTSD. Following the same procedure, we evaluate the feature importance based on the BIC score-based BN. The results are presented in Figure 6. We get fairly consistent results between the GL_2-regularized BN and the BIC score-based BN, especially providing the same seven important features. These results are consistent with what we got based on the Markov boundary analysis in Table 2.

6.3. Naïve Bayes Model

Finally, we construct a naïve Bayes (NB) model for predicting PTSD. NB is a probabilistic model that is based on the Bayes rule. It assumes that the features are mutually conditional independent given the class variable. In practice, because of its simplicity (small number of parameters), NB models predict well even when the assumption on which it is based is violated (Rish 2001). Additionally, NB exhibits considerable tolerance to missing data (Juhola and Laurikkala 2013).

In this study, we construct the NB model using the *e*1071 R-package (Meyer et al. 2019). According to the feature importance results for both GL_2-regularized BN and the BIC score-based BN, some features do not contribute

Figure 6. Feature Importance Box Plots Based on the BIC Score-Based BN Model



Note. The features are sorted by decreasing values of the median.

much to predicting PTSD. Adding those features may bias the model estimation and make the resulting model predict poorly. Accordingly, we construct the NB model with seven important features specified in our feature importance analysis.[11] Like the BIC score-based method, the NB model parameters were estimated using Laplace correction. We report the prediction performance of the resulting NB model in Table 3. The NB model performs the worst among the four BN models. However, it still improves the screening sensitivity of the currently used PC-PTSD-5 test by 0.2% and the specificity by 0.9%.

7. Summary and Conclusion

In this paper, we propose a BIC score-based BN, a GL_2-regularized BN using a new method based on group LASSO and a naïve Bayes BN. We build the GL_2-regularized BN to address four challenges: probabilistic classification, missing data, highly skewed features, and an extensive search space of network structure.

In our GL_2-regularized BN model, the conditional probability distribution of each node is defined using multinomial logistic regression. We use the group LASSO penalty, which yields a sparse model, making the model estimation more stable with missing data and highly skewed features. An ordering-based search algorithm with strong causality-based clinical knowledge is adopted to search for the network structure. As a result, our proposed GL_2-regularized BN-based model is highly competitive with the BIC score-based BN in predicting veterans' likelihood of suffering from PTSD. The GL_2-regularized BN and the BIC score-based BN predict better than naïve Bayes BN. One advantage of the GL_2-regularized BN over the BIC score-based BN is that the construction process yields some information about the relative importance of the features in predicting PTSD. We conjecture that our new method can also be used in other domains, but this is not the focus of our study. The task of exploring the strengths and weaknesses of the GL_2-regularized BN construction method in general domains remains to be done.

With the primary goal of predicting PTSD, the proposed BNs are generative models that estimate the joint distribution of all the observed variables. This joint distribution can then be used to estimate the probability of PTSD given the values of other variables. The advantages of this approach are it can easily make predictions if some variables are missing and provide a graph of the relationship between all variables. An alternative approach would be to directly model the conditional distribution of PTSD given the variables using a parametric method, such as (regularized) logistic regression or random forest. One challenge for these models is how to predict with missing data. Deriving a discriminative model that correctly addresses some of the difficulties in this data set, such as making predictions with missing values, would be an exciting extension of this work.

Our study contributes to the Veterans Health Administration in two ways. First, our proposed models could serve as a screening tool for identifying veterans with probable PTSD. Specifically, they improve the screening efficiency of an existing PTSD screening test, PC-PTSD-5, by incorporating veterans' demographic information, military service records, and comorbidity (TBI) diagnostic history. Second, we investigate how much the proposed models rely on each feature by considering the estimated feature importance. Feature importance provides a numerical measure of how informative each part is in diagnosing PTSD. The important features for predicting

PTSD are (in decreasing order of importance): PC-PTSD-5, service-connected flag, combat flag, agent orange flag, military sexual trauma flag, traumatic brain injury, and age.

Acknowledgments

This paper is adapted from Y. Tan's PhD dissertation (Tan 2020).

Endnotes

[1] A more thorough review of the literature about the latest research of PTSD using machine learning techniques is provided in Online Appendix A.

[2] The Markov boundary of a variable in a BN consists of its parents, its children, and other parents of its children. Given values of *all* variables in the Markov boundary of a variable, all other variables are irrelevant for the variable.

[3] A graphical description of conventional PTSD diagnostic and treatment referral process for a VA hospital is provided in Online Appendix B.

[4] A primer on BNs and causal models is given in Online Appendix C.

[5] A traditional rule of thumb is that 3 to 10 imputations typically suffice (Rubin 1987).

[6] Tabu list refers to the set of solutions that have been visited in the recent past.

[7] A graphical display of *MSE* score for the proposed GL_2-regularized BN model with prespecified sequence values of penalty parameter λ is provided in Online Appendix E.

[8] As our data are categorical, the change of sensitivity/specificity for the cutoff value of PTSD probability is not continuous. As a result, the real specificity of our model can be higher than 0.8.

[9] The increase of successfully detected PTSD veterans per year is calculated as $160{,}000 \times 23.09\% \times 3.72\% = 1{,}374$. In our sample, the proportion of PTSD-diagnosed veterans is 23.09%.

[10] The difference of successfully detected PTSD veterans per year between GL_2-regularized BN and BIC score-based BN is calculated as $160{,}000 \times 23.09\% \times 0.13\% = 48$.

[11] In unreported results, we experiment by constructing the NB model with all 14 predictors. The prediction performance gets significantly worse concerning all three metrics.

References

Adams RE, Boscarino JA (2005) Differences in mental health outcomes among Whites, African Americans, and Hispanics following a community disaster. *Psychiatry* 68(3):250–265.

Altmann A, Toloşi L, Sander O, Lengauer T (2010) Permutation importance: A corrected feature importance measure. *Bioinformatics* 26(10): 1340–1347.

Brewin CR, Andrews B, Valentine JD (2000) Meta-analysis of risk factors for posttraumatic stress disorder in trauma-exposed adults. *J. Consultant Clinical Psych.* 68(5):748–766.

Bryant R (2011) Post-traumatic stress disorder vs. traumatic brain injury. *Dialogues Clinical Neurosci.* 13(3):251–262.

Chen Q, Wang S (2013) Variable selection for multiply-imputed data with application to dioxin exposure study. *Statist. Medicine* 32(21): 3646–3659.

Cheng J, Greiner R, Kelly J, Bell D, Liu W (2002) Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence* 137(1–2):43–90.

Friedman N (1998) The Bayesian structural EM algorithm. Cooper GF, Moral S, eds. *Proc. 14th Annual Conf. on Uncertainty in Artificial Intelligence* (Morgan Kaufmann Publishers, San Francisco), 129–138.

FY VUP (2017) National Center for Veterans Analysis and Statistics, US Dept of Veterans Affairs, Washington, DC.

Galatzer-Levy IR, Karstoft KI, Statnikov A, Shalev AY (2014) Quantitative forecasting of PTSD from early trauma responses: A machine learning application. *J. Psychiatric Res.* 59(12):68–76.

Gaviria SL, Alarcón RD, Espinola M, Restrepo D, Lotero J, Berbesi DY, Sierra GM, et al. (2016) Socio-demographic patterns of posttraumatic stress disorder in Medellin, Colombia and the context of lifetime trauma exposure. *Disaster Health* 3(4):139–150.

Glover F (1989) Tabu search—Part I. *ORSA J. Comput.* 1(3):190–206.

Holeva V, Tarrier N (2001) Personality and peritraumatic dissociation in the prediction of PTSD in victims of road traffic accidents. *J. Psychosomatic Res.* 51(5):687–692.

Huang S, Li J, Ye J, Fleisher A, Chen K, Wu T, Reiman E (2012) A sparse structure learning algorithm for Gaussian Bayesian network identification from high-dimensional data. *IEEE Trans. Pattern Anal. Machine Intelligence* 35(6):1328–1342.

Juhola M, Laurikkala J (2013) Missing values: How many can they be to preserve classification reliability? *Artificial Intelligence Rev.* 40:231–245.

Kohavi R, Sahami M (1996) Error-based and entropy-based discretization of continuous features. Simoudis E, Han J, Fayyad U, eds. *Proc. 2nd Internat. Conf. on Knowledge Discovery and Data Mining* (AAAI Press, Palo Alto, CA), 114–119.

Koller D, Friedman N (2009) *Probabilistic Graphical Models: Principles and Techniques* (MIT Press, Cambridge, MA).

Lauritzen SL, Jensen F (2001) Stable local computation with conditional Gaussian distributions. *Statist. Comput.* 11(2):191–203.

McNally R, Mair P, Mugno B, Riemann B (2017a) Co-morbid obsessive-compulsive disorder and depression: A Bayesian network approach. *Psych. Medicine* 47(7):1204–1214.

McNally RJ, Heeren A, Robinaugh DJ (2017b) A Bayesian network analysis of posttraumatic stress disorder symptoms in adults reporting childhood sexual abuse. *Eur. J. Psychotraumatology* 8(suppl 3):1341276.

Meyer D, Dimitriadou E, Hornik K, Weingessel A, Leisch F, Chang CC, Lin CC (2017) e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien. *R package version 1.6–8*, 424.

Niblett T (1987) Constructing decision trees in noisy domains. Bratko I, Lavrač N, eds. *Proc. 2nd Eur. Working Session on Learn.* (Sigma, Bled, Yugoslavia), 67–78.

Norris FH, Slone LB (2013) Understanding research on the epidemiology of trauma and PTSD. *PTSD Res. Quart.* 24(2–3):1–13.

Omurca Sİ, Ekinci E (2015) An alternative evaluation of post traumatic stress disorder with machine learning methods. *Proc. Internat. Sympos. on Innovations in Intelligent Systems and Applications* (IEEE, New York), 1–7.

Penn M, Bhatnagar S, Kuy S, Lieberman S, Elnahal S, Clancy C, Shulkin D (2019) Comparison of wait times for new patients between the private sector and United States Department of Veterans Affairs medical centers. *JAMA Network Open* 2(1):e187096.

Prins A, Bovin MJ, Smolenski DJ, Marx BP, Kimerling R, Jenkins-Guarnieri MA, Kaloupek DG, et al. (2016) The primary care PTSD screen for DSM-5 (PC-PTSD-5): Development and evaluation within a veteran primary care sample. *J. General Internal Medicine* 31(10):1206–1211.

Rish I (2001) An empirical study of the naïve Bayes classifier. *Proc. IJCAI Workshop on Empirical Methods in Artificial Intelligence*, vol. 3 (Morgan Kaufmann Publishers, San Francisco), 41–46.

Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys* (John Wiley & Sons, New York).

Rubin DB, Schenker N (1991) Multiple imputation in health-are databases: An overview and some applications. *Statist. Medicine* 10(4):585–598.

Schalinski I, Teicher MH, Nischk D, Hinderer E, Müller O, Rockstroh B (2016) Type and timing of adverse childhood experiences differentially affect severity of PTSD, dissociative and depressive symptoms in adult inpatients. *BMC Psychiatry* 16(1):295.

Schwarz G (1978) Estimating the dimension of a model. *Ann. Statist.* 6(2):461–464.

Scutari M (2010) Learning Bayesian networks with the *bnlearn R* package. *J. Statist. Software* 35(3):1–22.

Scutari M, Graafland CE, Gutiérrez JM (2018) Who learns better Bayesian network structures: Constraint-based, score-based or hybrid algorithms? Kratochvíl V, Studený M, eds. *Proc. Internat. Conf. on Probabilistic Graphical Models* (PMLR, New York) 416–427.

Spirtes P, Glymour C (1991) An algorithm for fast recovery of sparse causal graphs. *Soc. Sci. Comput. Rev.* 9(1):62–72.

Sun Y, Tang Y, Ding S, Lv S, Cui Y (2011) Diagnose the mild cognitive impairment by constructing Bayesian network with missing data. *Expert Systems Appl.* 38(1):442–449.

Tan Y (2020) New probabilistic techniques for classification problems and an application. PhD thesis, University of Kansas School of Business, Lawrence, KS.

Tan Y, Shenoy PP, Sherwood B, Gaddy M, Oehlert ME, Shenoy C (2023) R scripts for Bayesian network models for PTSD screening in veterans. https://dx.doi.org/10.1287/ijoc.2021.0174.cd, https://github.com/INFORMSJoC/2021.0174.

Teyssier M, Koller D (2005) Ordering-based search: A simple and effective algorithm for learning Bayesian networks. Bacchus F, Jaakkola T, eds. Uncertainty in Artificial Intelligence: Proc. *21st Conf.* (AUAI Press, Arlington, VA), 584–590.

Tibshirani R (1996) Regression shrinkage and selection via the lasso. *J. Royal Statist. Soc. B* 58(1):267–288.

Tsamardinos I, Brown LE, Aliferis CF (2006) The max-min hill-climbing Bayesian network structure learning algorithm. *Machine Learn.* 65(1):31–78.

U.S. Department of Veterans Affairs (2014) 172VA10P2: VHA corporate data warehouse. *Federal Register* 79(17):4377–4382.

Van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Statist. Methods Medical Res.* 16(3):219–242.

Vincent M, Hansen NR (2014) Sparse group lasso and high dimensional multinomial classification. *Comput. Statist. Data Anal.* 71:771–786.

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. *J. Royal Statist. Soc. B* 68(1):49–67.

Zhu J, Hastie T (2004) Classification of gene microarrays by penalized logistic regression. *Biostatistics* 5(3):427–443.