HHS Public Access

Author manuscript

J Biomed Inform. Author manuscript; available in PMC 2018 November 01.

Published in final edited form as:

J Biomed Inform. 2017 November ; 75 Suppl: S94–S104. doi:10.1016/j.jbi.2017.05.019.

Predictive Modeling for Classification of Positive Valence System Symptom Severity from Initial Psychiatric Evaluation Records

Jose D. Posada[a,b,*], **Amie J. Barda**[a,*], **Lingyun Shi**[a], **Diyang Xue**[a], **Victor Ruiz**[a], **Pei-Han Kuan**[c], **Neal D. Ryan**[d], and **Fuchiang (Rich) Tsui**[a,**]

[a]Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Blvd., Pittsburgh, PA 15206

[b]Electronics and Telecommunications Engineer Program, Universidad Autónoma del Caribe, CI. 90 #46-112, Barranquilla, Atlántico, Colombia

[c]Institute of Manufacturing Information and System, National Cheng-Kung University, Tainan, Taiwan

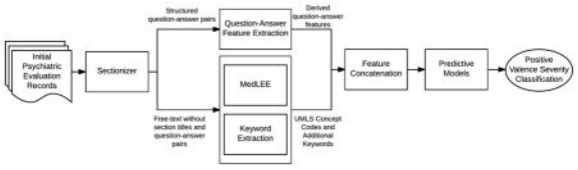
[d]Department of Psychiatry, University of Pittsburgh, 3811 O'Hara St., Pittsburgh, PA 15213

Abstract

In response to the challenges set forth by the *CEGS N-GRID 2016 Shared Task in Clinical Natural Language Processing*, we describe a framework to automatically classify initial psychiatric evaluation records to one of four positive valence system severities: absent, mild, moderate, or severe. We used a dataset provided by the event organizers to develop a framework comprised of natural language processing (NLP) modules and 3 predictive models (two decision tree models and one Bayesian network model) used in the competition. We also developed two additional predictive models for comparison purpose. To evaluate our framework, we employed a blind test dataset provided by the 2016 CEGS N-GRID. The predictive scores, measured by the macro averaged-inverse normalized mean absolute error score, from the two decision trees and Naïve Bayes models were 82.56%, 82.18%, and 80.56%, respectively. The proposed framework in this paper can potentially be applied to other predictive tasks for processing initial psychiatric evaluation records, such as predicting 30-day psychiatric readmissions.

Graphical Abstract

**Corresponding author. *Phone:* 412-648-7182. tsui2@pitt.edu.
*These authors contributed equally to this work as first authors.



1. Introduction

The *CEGS N-GRID 2016 Shared Task in Clinical Natural Language Processing* put forth three competition challenge tracks for a corpus of 816 initial psychiatric evaluation records: De-identification (Track 1) [1], Research Domain Criteria (RDoC) classification (Track 2) [2], and novel data use to investigate questions beyond those posed by the challenge organizers (Track 3). In this paper, we describe a framework to address the Track 2 challenge of classifying initial narrative psychiatric evaluation records per the RDoC framework [3].

In contrast to current categorical diagnostic systems (*e.g.,* DSM-5, ICD-10), the RDoC framework attempts to classify mental disorders based on “dimensions of observable behavior and neurobiological measures”, with the goal of stimulating new approaches to mental disorder research[4]. The main framework is divided into five psychiatric domains of functioning: positive valence systems (PVS), negative valence systems (NVS), cognitive systems, systems for social processes, and arousal/regulatory systems [5]. Each domain consists of a set of functional constructs (*i.e.,* concepts representing a specified functional dimension of behavior such as approach motivation) that are characterized at different levels (*e.g.,* genomic, molecular, cellular, circuital, physiological, behavioral, self-reported or paradigmatic) [3]. Reliably classifying symptom severity within the five RDoC domains is critical to implementing and validating the RDoC approach [4].

The Track 2 challenge focused on classifying initial psychiatric evaluation records by symptom severity within an RDoC domain using an ordinal severity scale from 0 to 3: absent (0), mild (1), moderate (2), severe (3). The challenge focused specifically on the PVS domain, which spans those brain systems and related feelings and behaviors involved in contexts including reward seeking, enjoying pleasurable experiences, and habit learning. These systems are thought to play an important role in the initiation and maintenance of many psychiatric disorders including substance (*e.g.,* opioid) use disorders, major depressive disorder, and bipolar disorder[4].

We propose a largely automated framework comprised of data processing and predictive models to address the challenge of classifying initial psychiatric evaluation records by symptom severity within the PVS domain. We hypothesize that the proposed framework can be used to accurately classify individual initial psychiatric evaluation records into one of four severity levels within an RDoC domain. In line with the proposed Task 2 challenge, we developed and tested the framework using the PVS domain, but believe that it could be applied to the other RDoC domains.

2. Methods

2.1 Data

The *CEGS N-GRID 2016 Shared Task in Clinical Natural Language Processing* allowed challenge participants access to a corpus of 816 initial psychiatric evaluation records provided by Partners Healthcare and the N-GRID project of Harvard Medical School. All records were fully de-identified by the challenge organizers prior to distributing to participants. For the Task 2 challenge, the event organizers released data in two stages. The initial stage included 600 (433 annotated and 167 unannotated) initial psychiatric evaluation records (in XML format). Each annotated record was assigned a single PVS symptom severity classification on an ordinal scale from 0 to 3 as follows [2]:

0. **Absent:** no symptoms

1. **Mild:** some symptoms present but not a focus of treatment

2. **Moderate:** symptoms present and a focus of treatment but not requiring hospitalization or equivalent

3. **Severe:** symptoms present requiring hospitalization, emergency department visit, or otherwise having a major consequence

Of the 433 initial annotated records, 325 were annotated by two psychiatrists and 108 records were annotated by only one psychiatrist. These 433 annotated records comprised our training dataset. We did not make use of the initial unannotated records in training or testing because our framework relies on supervised classification algorithms. In the second stage, 216 unannotated records were provided to participants 3 days prior to the competition deadline. These records comprised our blind test dataset. After the deadline, annotations from those 216 records were released to the participants to self-evaluate our performance.

2.2. Framework

Figure 1 summarizes our proposed framework, which contained multiple natural language processing (NLP) components and 3 predictive models.

2.2.1 Sectionizer Component—Two team members (JP, LS) identified a set of 238 section titles (*e.g.,* Chief Complaint) and structured question-answer pairs (*e.g.,* Hx[1] of Suicidal Behavior: Yes) through an iterative manual review of the records in the training dataset. The sectionizer component, written in Java code, extracted a set of section titles and question-answer pairs from the records. The sectionizer then removed Section titles and the set of structured questions from the records to reduce the chance of detecting false positives during NLP processing, but answers from the structured question-answer pairs were not removed. The complete set of extracted question-answer pairs were stored separately for further processing. The sectionized records (*i.e.,* records with section titles and structured questions removed) were then passed to the MedLEE and Keyword Extraction components.

[1]Hx: History

2.2.2 MedLEE and Keyword Extraction Components—The MedLEE component processed the sectionized records using MedLEE [6], a clinical NLP tool that identifies clinical terms and outputs the corresponding Unified Medical Language System (UMLS) codes. Although MedLEE identifies a large proportion of important clinical terms, it has not been adapted to specifically address the domain of psychiatry and therefore misses important clinical indications in the field, such as domain-specific abbreviations and social factors. We developed the Keyword Extraction component to address these gaps. We randomly selected twenty records (five records from each of the four severity levels) to develop this component. We used the list of UMLS codes identified by MedLEE and manually reviewed the 20 sampled records to develop a list of missing terms, phrases, and abbreviations that were deemed potentially relevant to the RDoC severity score classification problem. Collectively we called the identified terms, phrases, and abbreviations the 'keyword list'. The keyword list contained many nonstandard abbreviations used in psychiatry (*e.g.,* SI, which is an acronym for suicidal ideation) and social factors important in RDoC severity classification (*e.g.,* arrests, probation, homeless, unemployment, lost custody of child, etc.). To increase retrieval of keywords, all single terms were reduced to their base form (*e.g.* "arrested" is reduced to "arrest") using lemmatization in Standford's CoreNLP[7]. The Keyword Extraction component used the keyword list to process the records and extract additional information not identified by MedLEE. The extracted keywords were then grouped into nine categories (consequences (any), hospitalization, legal consequences, social consequences, substance abuse, consequences due to substance abuse, treatment of substance abuse, suicidal/self-harm, treatment (any)) and counts of keywords found within each category were extracted as features. We also included the counts of individual keywords that could not be grouped into a category (*e.g.,* PTSD, which represents a common acronym for post traumatic stress disorder. We combined the Keyword Extraction component output with the MedLEE component output that was converted to binary values (*i.e.,* presence or absence) and passed the combined output to the Feature Concatenation component.

2.2.3 Question-Answer Feature Extraction and Feature Concatenation Components—The Sectionizer component extracted a set of 124 structured question-answer pairs from the records. The Question-Answer Feature Extraction component processed this set to generate features. First, the set of 124 questions was reviewed by a psychiatrist (NR) to identify PVS relevant questions. A second team member (AB) reviewed the set to identify any questions potentially relevant to severity classification. This resulted in 61 potentially relevant questions identified. For each identified question, a set of predefined set of answers was generated. Most questions were categorical in nature and could be classified using 'yes', 'no', 'missing', or 'uncertain', although a few questions required individualized answer sets. For example, smoking status of patients was defined using categories of 'current', 'former', 'never', and 'missing'. All observed answers for each categorical question were then standardized by mapping to the pre-defined answer sets. For score-type questions (*e.g.,* Audit-C score), numeric information was extracted. We grouped extracted ICD-9 codes to a family (integer) level (*e.g.,* 300.XX) and converted them to categorical values (*i.e.*, yes, no, and missing). Diagnostic and Statistical Manual of Mental Disorder (DSM) axis IV codes were extracted, mapped to the nine defined axis IV

categories, and converted to categorical values (*i.e.,* yes, no, and missing). We then removed questions with missing answers across most records. Finally, we derived eleven new 'score' features through aggregated counts of the answers to related or similar questions. For example, we generated a depression 'score' feature by identifying two questions related to depression and summing the number of positive ('yes') answers to those questions (*i.e.,* the possible depression 'score' values ranged from 0 to 2).

The Feature Concatenation module concatenated features generated by the Question-Answer Extraction, MedLEE, and Keyword Extraction components. We then passed the final concatenated feature set that can be found in Appendix 1 to three predictive models.

2.2.4 Predictive Models—We used all 433 annotated records to build three predictive models for the competition: two decision tree (DT) models and one Bayesian network (BN) model. We also added an additional hierarchy-based BN model and a commonly used baseline model: a support vector machine (SVM) model applied to a term frequency–inverse document frequency (TF-IDF) matrix without the use of our proposed pipeline.

We employed the "rpartScore" package in R to build the decision tree (DT) models[8]. The package provides functions to build classification trees for ordinal responses within the classification and regression tree (CART) framework. This process involves two phases: splitting and pruning. During the split phase, trees are grown utilizing a recursive partitioning procedure wherein a node and binary partition are selected to minimize node impurity as measured by the generalized Gini impurity function. For a set of items with J classes, the generalized Gini impurity function for a node t is defined as:

$$i_{GG}(t)=\sum_{k=1}^{J}\sum_{l=1}^{J}C(\omega_k|\omega_l)p(\omega_k|t)p(\omega_l|t) \quad (1)$$

where $p(\omega_k|t)$ is the proportion of items in node t belonging to the k-th category and $C_{SD}(\omega_k|\omega_t)$ and $C_{Ab}(\omega_k|\omega_t)$ are misclassification costs of assigning category ω_k to an item actually belonging to category ω_l. Misclassification costs are computed using either the squared difference in scores of the absolute difference in scores as defined in Equations 2 and 3, respectively,

$$C_{SD}(\omega_k|\omega_l)=(s_k-s_l)^2, \quad (2)$$

$$C_{Ab}(\omega_k|\omega_l)=|s_k-s_l| \quad (3)$$

where s_k is the score for category k. The trees produced in the splitting phase are then pruned to avoid overfitting. Pruning can be based on the total misclassification rate ($R_{mr}(T)$) or the total misclassification cost ($R_{mc}(T)$) as defined in equations 4 and 5, respectively,

$$R_{mr}(T)=\sum_{i=1}^{n}[1-I_{\{s_i\}}(\hat{s}_{i,T})] \quad (4)$$

$$R_{mc}(T)=\sum_{i=1}^{n}|s_i-\hat{s}_{i,T}| \quad (5)$$

where s_i is the observed score for item i, $\hat{s}_{i,T}$ is the predicted score for item i by tree T, and $I_{\{s_i\}}(\hat{s}_{i,T}) = 1$ if $s_i = \hat{s}_{i,T}$ and 0 otherwise.

Both DT models were built using three hierarchical steps: (1) classify patients into “absent/mild” or “moderate/severe” severity subgroup, (2) classify patients in the “absent/mild” subgroup as “absent” or “mild”, and (3) classify patients in the “moderate/severe” subgroup as “moderate” or “severe”. The two DT models differed in the misclassification calculations used during the splitting and pruning phases. The first DT model used squared difference in scores as misclassification cost C_{SD} (Equation 2) in the splitting phase and total misclassification cost R_{mc} in the pruning phase (Equation 5). The second DT model used absolute difference in scores as misclassification cost C_{Ab} (Equation 3) in the splitting phase and total misclassification rate R_{mr} in the pruning phase (Equation 4). These two models are referred to as the “DT SD-MC” model and the “DT Ab-MR” model, respectively.

Our third predictive model was an Ordinal-response Multiple Bayesian Networks (OMBN) model. To build the network, we first performed correlation-based feature selection (CFS) [9], which aims to find a set of features that are highly correlated with the prediction class, yet are uncorrelated with each other. This is accomplished by assigning a heuristic merit score (M_S) to each feature subset S consisting of k features, defined as

$$M_S=\frac{k\overline{r_{cf}}}{\sqrt{k+k(k-1)\overline{r_{ff}}}} \quad (6)$$

where $\overline{r_{cf}}$ is the average value of all class-feature correlations and $\overline{r_{ff}}$ is the average value of all feature-feature correlations. CFS aims to find the feature subset that maximizes the heuristic merit score criterion defined by Equation 6. After feature selection, we trained a Bayesian network classifier with structure learned from the K2 search algorithm described by Cooper and Herskovits[10]. Classification of ordinal responses using a traditional Bayesian model was achieved following the approach described by Eibe and Hall [11]. Appendix 2 shows the OMBN modeling process diagram and corresponding Bayesian networks.

The fourth predictive model that was not included in the competition was a Hierarchy-based Multiple Bayesian Networks (HMBN) model. It followed the three hierarchical steps described in the aforementioned DT modeling to build three BNs following CFS feature selection. The first BN was trained and classified among two subgroups of patients: “absent/

mild" or "moderate/severe" severity subgroups. The second BN was trained and classified among two granular severity subgroups: "absent" or "mild", and the third BN was trained and classified among two another granular severity subgroups: "moderate" or "severe".

For a baseline comparison (not included in the competition), we trained a linear support vector machine (SVM) and preprocessed the text without making use of our pipeline. The preprocessing was performed using the Python package NLTK [12]. The preprocessing included standard processes such as: sentence boundary detection, tokenization, stop word and punctuation removal, and lemmatization. After preprocessing the text, a term frequency–inverse document frequency (TF-IDF) matrix was created to train four linear SVMs using the Python package scikit-learn [13]. We used a one vs. all approach to achieve multiclass classification.

2.3 Model Evaluation

The gold standard of positive valence system severity for each set of records was determined by psychiatrists who read and annotated the records. In accordance with the challenge guidelines, we evaluated classification performance of the predictive models on the test (blind) dataset against the gold standard using the macro averaged-inverse normalized mean absolute error ($INMAE^M$) score. A detailed description of the $INMAE^M$ score is available elsewhere[2]. The scoreranges from 0 to 100, where 100 indicates the highest performance. For each predictive model, an overall $INMAE^M$ score was calculated across the combined severity classes. $INMAE^M$ scores were also calculated for each individual severity class.

2.3 Data Ablation Evaluation

We identified three types of features in this study as shown in Figure 1: Question-Answer pairs (Q&A), MedLEE extracted features, and Keyword extraction (Keywords). To better understand the contribution of individual feature types and any combinations of the three feature types for prediction performance, we employed the best model approach identified from the test-data performance to the various combinations of the three feature types: Q&A, MedLEE, Keywords, Q&A + MedLEE, Q&A + Keywords, and MedLEE + Keywords.

3. Results

Using the framework outlined in Figure 1, we obtained total 5,447 final features in the training dataset comprised of 5,330 unique UMLS concepts from the MedLEE component, 16 features from the Keyword Extraction component, and 101 features from the Question-Answer Feature Extraction component. 22.34% of questions had a missing value, and 77.66% contained an answer. Appendix 1 lists all features and observed values used by the predictive models.

3.1 Predictive Models

We tested five predictive models: DT SD-MC, DT Ab-MR, OMBN, HMBN, and a linear SVM. Both HMBN and the SVM were developed after the competition for the purpose of comparison. The linear SVM with the standard parameters on the scikit-learn package served as a baseline modele competition. The best performing model, DT SD-MC,

comprised three DTs as shown in Figures 2, 3 and 4. The final OMBN is depicted in Appendix 2. A total of 25 features were included in the trees with best performance. From those 25 features, 16 (64%) were extracted using the question-answer feature extraction component. Figures 2 to 4 show that many of the discriminating features are associated with some form of substance abuse.

Table 1 summarizes the final evaluation of our five predictive models. The evaluation was done using $INMAE^M$ scores. For the best performing model (DT SD-MC) the score difference between the test set and training set score is minimal. In almost all cases the best performance was observed for classifying absent symptom severity. The OMBN had a consistently lower performance for almost all the classes except for the moderate class. On the other hand, the HMBN had the best performance on the training set. Despite HMBN exhibiting the best performance, it was not trained during the competition and was therefore not submitted as our best model. All ls had a better overall performance than the baseline SVM model on both the training and testing data sets.

Table 2 shows an ablation study where the contribution of each set of features is assessed individually using the test set. For three of the four classes, features from the question-answer pairs were present on the best performing models. Performance from the combination of question-answer pairs with MedLEE features was consistently the best or the second best in all experiments.

4. Discussion

We found that our proposed framework exhibited promising performance when classifying symptom severity within the PVS domain. Prior to this challenge event, there has been limited work in classifying symptom severity according to the RDoC framework and no prior work using NLP techniques to tackle the problem. The development of reliable and valid severity coding of RDoC domains using only textual data has the potential to expend RDoC to large naturalistic datasets. Despite some limitations in our proposed framework, much of it is automated and reusable. With minor adaptations of the framework (*e.g.*, Keyword Extraction component), it could be applied to other predictive tasks related to psychiatric records.

4.1 Significance of findings

We successfully classified PVS symptom severity within free-text initial psychiatric evaluation records using NLP techniques and predictive modeling. Outside of this challenge event, no other work has been done that uses NLP techniques to classify symptom severity as per the RDoC framework. Others have utilized other techniques to classify symptom severity, such as external assessment scales [14], but to our knowledge no previous work has attempted to assess symptom severity using routinely collected free-text reports.

Our best performing model (DT SD-MC) exhibited minimal score difference between the test set and training set. This minimal difference could be attributed to good generalization capacity for the model. The difference between the DT SD-MC and DT Ab-MR models could be attributed to differences in performance among moderate and severe classes, as

observed in Table 1. The lower performance of the OMBN modeling approach could be attributed to the subtle differences between some of the severity levels that were better accounted for in the hierarchical modeling approach of the DT models. For example, differentiating between absent/mild and moderate/severe was easily identified in the DT models by singular factors such as substance abuse. On the other hand, distinguishing between moderate and severe relied on identifying more subtle factors of a disorder, such as social consequences. By splitting up the classification tasks hierarchically, *i.e.,* classifying patients into absent/mild and moderate/severe groups and then classifying each subgroup (e.g., absent vs. mild), the models may have been able to better detect subtle class differences. By not manually defining group hierarchy, the ability of the OMBN modeling approach to detect subtle class differences may have been diminished, thus resulting in poorer performance when compared to the DT approaches. In addition, even after using the same hierarchical strategy with the Bayesian networks, HMBN had a lower performance on the test set than the DT SD-MC, indicating that it had a poorer generalization capacity.

In addition to better performance, the DT approaches allow for straightforward interpretation of the classification process that is not available with the other approaches. One may argue that the misclassification from the first tree propagates to the others because the three decision trees could be seen as a single model, where each leaf on Figure 2 is replaced with the correspondent tree from Figure 3 or 4. In other words, the two “A or MI” leaf nodes in the tree in Figure 2 could be replaced with the first node in the tree in Figure 3 and the seven “M or S” leaf nodes in Figure 2 could be replaced with the first node in the tree in Figure 4. The result would be a singular decision tree where each leaf node represents only one of the four severity classifications. Despite the potential for error propagation, the value of the single decision tree lies in its straightforward interpretation that could easily translate into a clinical decision making process.

The importance of correctly extracting the information from structured questions in the record is reflected in that 64% of all features present in the best performing model came from the question-answer feature extraction component. Also, results from Table 2 showed that question-answer pairs features played an important role in the overall performance of the system when compared against all the other features. Among the four best performing experiments in Table 2, question-answer pairs are present in three of them. These results could support the hypothesis that most of the information in initial psychiatric evaluation records is encoded in these question-answer pairs rather than the free-text sections. A more formal investigation of this matter is required.

Furthermore, many of the features selected by our models are clinically relevant, as validated by the literature. Our models picked up features associated with disorders that have symptoms attributable to disruptions within PVS constructs, such as bipolar disorders, obsessive compulsive disorders (OCD), anxiety disorders [15], depressive disorders [14,16], eating disorders [17], and substance abuse disorders [16,18]. In particular, features associated with substance abuse were quite prevalent in our models, where positive values for the features tended to result in higher symptom severity classifications. This was a particularly encouraging finding as substance abuse and addiction disorders are known to be strongly associated with disruptions in PVS constructs[16,18,19]. These findings provide

evidence that our methods could be potentially useful in diagnosing and treating disorders with symptoms associated with PVS construct disruptions.

4.2 Limitations

Given that the number of training samples were fewer than the total number of initial features, a possibility for overfitting is present. Considering, however, that the larger, better performing DT model had only 8 variables, the possibility of overfitting is less concerning. On the other hand, using the identified features as risk factors or in an automated model to assign a severity score to a patient is limited to the current sample size and the singular hospital location. The sample size and singular site limits generalizability of the results for the entire population and prevents extrapolation of the results to other hospitals. Furthermore, difference in clinical practice could result in vast differences in the sections and structured question-answer pairs found in a report. As we manually reviewed the reports to identify the list of possible sections and question-answer pairs, this portion of our work would have to be redone if reports from other psychiatric settings are radically different from those in our dataset. Given the proposed framework, however, adaptations that respond to such challenges could be easily implemented. The implementation efforts would focus on the manual identification of sections and questions, a task that is supported by the described Sectionizer and can be completed in a reasonable amount of time. Finally, MedLEE is a commercial software which needs to be purchased to be used. This limitation could be alleviated by replacing the MedLEE component with an open source biomedical NLP tool such as cTakes[20].

4.3 Potential applications of framework

Although we developed our framework centering around the PVS domain, the framework could be applied to classify symptom severity within other RDoC domains with minimal adaptations. More generally, there are numerous problems within the psychiatric domain for which text mining approaches have been applied [21]. Our NLP methods could be adapted to further explore these problems and potentially expand upon the current proposed solutions. For example, it has been shown that NLP can aid in identifying psychiatric patients at risk of early readmission from narrative discharge summaries [22]. By modifying the Sectionizer component to handle narrative discharge summaries and removing the PVS filtering in the Question-Answer-Extraction component, we could apply our methods to the task of predicting psychiatric readmission from narrative discharge summaries. This may result in improved performance and better understanding of the readmission problem as our methods can account for more information than the topic modeling methods originally applied. As our framework is replicable with minimal time requirement, it could potentially be adapted to any prediction problem that utilizes free-text psychiatric records.

5. Conclusion

In response to the challenges outlined in the CEGS N-GRID 2016 Shared Task in Clinical Natural *Language Processing*, we developed a framework for processing and classifying symptom severity within initial psychiatric evaluation records according to the RDoC framework. Our proposed framework exhibited promising classification performance within

the PVS domain and can be easily adapted to address other predictive tasks related to free-text psychiatric records, such as symptom severity classification within the other RDoC domains and improving prediction of 30-day psychiatric readmission.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the Coulter foundation, a local charity foundation in Pittsburgh, the Richard King Mellon foundation, the National Library of Medicine award R01 LM011370, the National Library of Medicine 5 T15 LM007059 to the University of Pittsburgh's Biomedical Informatics Training Program and Fulbright Grant for the development regions. The organization of the *CEGS N-GRID 2016 Shared Task in Clinical Natural Language Processing* was made possible through the support of the National Institutes of Health [NIH P50 MH106933 (PI: Isaac Kohane); NIH 4R13LM011411 (PI: Ozlem Uzuner).

References

1. Stubbs A, Filannino M, Uzuner Ö. De-identification of psychiatric intake records: Overview of 2016 CEGS N-GRID Shared Tasks Track 1. J Biomed Inform. 2017
2. Filannino M, Stubbs A, Uzuner Ö. Symptom severity prediction from neuropsychiatric clinical records: Overview of 2016 CEGS N-GRID Shared Tasks Track 2. J Biomed Inform. 2017
3. Cuthbert BN, Insel TR. Toward the future of psychiatric diagnosis: the seven pillars of RDoC. BMC Med. 2013; 11:126.doi: 10.1186/1741-7015-11-126 [PubMed: 23672542]
4. NIMH. Positive Valence Systems: Workshop Proceedings. Res Domain Criteria. 2011. http://www.nimh.nih.gov/research-priorities/rdoc/n
5. Kozak MJ, Cuthbert BN. The NIMH Research Domain Criteria Initiative: Background, Issues, and Pragmatics. Psychophysiology. 2016; 53:286–97. [accessed November 3, 2016] http://www.ncbi.nlm.nih.gov/pubmed/26877115. [PubMed: 26877115]
6. Friedman, C. A broad-coverage natural language processing system; AMIA Annu Symp Proc. 2000. p. 270-4.http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2243979&tool=pmcentrez&rendertype=abstract
7. Manning, CD., Bauer, J., Finkel, J., Bethard, SJ., Surdeanu, M., McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. Proc. 52nd Annu. Meet. Assoc. Comput. Linguist. Syst. Demonstr; 2014. p. 55-60.
8. Galimberti G, Soffritti G, Di Maso M. Classification Trees for Ordinal Responses in R : The rpartScore Package. J Stat Softw. 2012; 47:25. http://dx.doi.org/10.18637/jss.v047.i10.
9. Hall M. Correlation-based Feature Selection for Machine Learning. Methodology. 1999; 21i195-i20:1–5. doi:10.1.1.149.3848.
10. Cooper GF, Herskovits E. A bayesian method for the induction of probabilistic networks from data. Mach Learn. 1992; 9:309–347. DOI: 10.1007/BF00994110
11. Frank E, Hall M. A simple approach to ordinal classification. Mach Learn ECML 2001. 2001; 2167:145–156. DOI: 10.1007/3-540-44795-4_13
12. Bird, S., Klein, E., Loper, E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc; 2009.
13. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay É. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2012; 12:2825–2830. DOI: 10.1007/s13398-014-0173-7.2
14. Toups M, Carmody T, Greer T, Rethorst C, Grannemann B, Trivedi MH. Exercise is an effective treatment for positive valence symptoms in major depression. J Affect Disord. 2016; 209:188–194. DOI: 10.1016/j.jad.2016.08.058 [PubMed: 27936452]

15. Bress JN, Meyer A, Hajcak G. Differentiating anxiety and depression in children and adolescents: evidence from event-related brain potentials. J Clin Child Adolesc Psychol. 2015; 44:238–249. DOI: 10.1080/15374416.2013.814544 [PubMed: 23879474]
16. Baskin-Sommers AR, Foti D. Abnormal reward functioning across substance use disorders and major depressive disorder: Considering reward as a transdiagnostic mechanism. Int J Psychophysiol. 2014; 98:227–239. DOI: 10.1016/j.ijpsycho.2015.01.011
17. Wildes JE, Marcus MD. Application of the Research Domain Criteria (RDoC) Framework to Eating Disorders: Emerging Concepts and Research. Curr Psychiatry Rep. 2015; 17doi: 10.1007/s11920-015-0572-2
18. Sánchez E, Cruz-Fuentes C. Cognitive Control and Negative and Positive Valence Systems in the Development of an NIMH RDoC-Based Model for Alcohol Use Disorder. Alcohol Clin Exp Res. 2016; 40:214–5. [accessed January 26, 2017] http://www.ncbi.nlm.nih.gov/pubmed/26727536. [PubMed: 26727536]
19. Schneider S, Peters J, Bromberg U, Brassen S, Miedl SF, Banaschewski T, Barker GJ, Conrod P, Flor H, Garavan H, Heinz A, Ittermann B, Lathrop M, Loth E, Mann K, Martinot J-L, Nees F, Paus T, Rietschel M, Robbins TW, Smolka MN, Spanagel R, Ströhle A, Struve M, Schumann G, Büchel C. Risk Taking and the Adolescent Reward System: A Potential Common Link to Substance Abuse. Am J Psychiatry. 2012; 169:39–46. [PubMed: 21955931]
20. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, Chute CG. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Informatics Assoc. 2010; 17:507–513. http://jamia.oxfordjournals.org/content/17/5/507.abstract.
21. Abbe A, Grouin C, Zweigenbaum P, Falissard B. Text mining applications in psychiatry: a systematic literature review. Int J Methods Psychiatr Res. 2015; n/a-n/a. doi: 10.1002/mpr.1481
22. Rumshisky A, Ghassemi M, Naumann T, Szolovits P, Castro VM, McCoy TH, Perlis RH. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Transl Psychiatry. 2016; 6:e921.doi: 10.1038/tp.2015.182 [PubMed: 27754482]

Appendix 1. Description of feature set

Name	Description	Component	Observed Values
UMLS Codes	Indicated by format of C#######. Included 5330 codes as features	MedLEE	{Present, Absent}
Bipolar	Category of keywords indicating bipolar disorder (e.g., BP, BPAD); expressed as count of terms in category	Keyword Extraction	{0, 1, 2, 3, 4, 5, 7}
BPD	Identified keyword; abbreviation for borderline personality disorder; expressed as count		{0, 1, 3}
consequence_binary_v1	Category of keywords associated with some sort of consequence, whether legal, social or otherwise; expressed as count		{0, 1}
consequence_count	Category of keywords associated with some sort of consequence, whether legal, social or otherwise; expressed as count		{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 16, 18, 19, 20, 26, 28}
detain	Identified keyword; expressed as count		{0, 1, 2}

Name	Description	Component	Observed Values
HI	Identified keyword; abbreviation for homicidal ideation; expressed as count		{0, 1, 2, 3, 4, 5}
Hospitalization	Category of keywords indicating hospitalization (e.g., hospd, EMS); expressed as count of terms in category		{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11}
Legal_consequences	Category of keywords associated with legal consequences (e.g., restraining order); expressed as count of terms in category		{0, 1, 2, 3, 4, 5, 6, 7, 8, 15, 17, 21}
PTSD	Identified keyword; abbreviation for post-traumatic stress disorder; expressed as count		{0, 1, 2, 3, 4, 5, 6, 7, 8, 10, 11, 12}
Rehabilitate	Identified keyword; expressed as count		{0, 1}
Social_consequences	Category of keywords associated with social consequences (e.g., lost custody of child); expressed as binary variable		{Present, Absent}
Substance_consequence_overlap	Category of keywords associated with consequence related to substance abuse (e.g., DUI, FAS); expressed as count of terms in category		{0, 1, 2, 3, 4}
Substance_related	Category of keywords associated with substance abuse (e.g., o.d., EtOH, psychedelic drug); expressed as count of terms in category		{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 22, 24, 28, 30, 31, 35}
Substance_treatment_overlap	Cateogry of keywords associated with treatment of a substance abuse problem (e.g., sobriety, AA meeting); expressed as count of terms in category		{0, 1, 2, 3, 5, 6, 7, 8, 9, 10, 11, 12, 13, 15, 17, 18, 19}
Suicidal_selfHarm	Category of keywords indicating suicidal or self-harm tendencies (e.g., SI, SA); expressed as count of terms in category		{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 14}
Treatment_related	Category of keywords associated with treatment (e.g., tx, CBT/cognitive behavioral therapy); expressed as count of terms in category		{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15 16, 17, 18, 19, 20, 21, 23, 27, 28, 35}
Q91	"How many packs of cigarettes per day does the patient smoke?"	Question- Answer Feature Extraction	{<=1, >1, Missing}
Q_BIPOLAR	Composite feature; sum score of positive answers to 2 bipolar disorder assessment questions		{0, 1, 2}
Q_DEMENTIA	Composite feature; sum score of positive answers to 2 dementia assessment questions		

Name	Description	Component	Observed Values
Q_DEPRESS	Composite feature; sum score of positive answers to 2 depression assessment questions		
Q_EATING	Composite feature; sum score of positive answers to 2 eating disorder assessment questions		
Q_OCD	Composite feature; sum score of positive answers to 2 OCD assessment questions		
Q_PSYCH	Composite feature; sum score of positive answers to 2 psychosis assessment questions		
Q_PTSD	Composite feature; sum score of positive answers to 2 PTSD assessment questions		
Q_ALCOHOL	Composite feature; total score on 3 alcohol use questions where score on each question ranges from 0 to 4 (i.e. audit c score calculation)		
Q_AUDITC	Composite feature; maximum recorded score from 4 questions on Audit C Score (i.e., current score, score at clinic intake date, highest recorded score to date, total score)		{0, 1,…,12}
Q_SUICIDE	Composite feature; sum score of positive answers to 2 suicide assessment questions		{0, 1}
Q23	CPT codes recorded by psychiatrist		{90791, 90792, 90801, Missing}
Q72	Assessment of judgement		{Abnormal, WNL, Missing}
Q109	"What is the smoking status of the patient?"		{Current, Former, Never, Missing}
Q117	"Does the patient frequently make movements they can't control?" (Tourette's assessment)		
Q118	"Does the patient frequently make noises they can't control?" (Tourette's asssessment)		{No, Missing}
Q2	"Are there any additional risk issues related to the illness/ treatment assessed today?"		
Q50	"Is there a family history of suicidal behavior?"		
Q107	"Does the patient have any thoughts about self abuse?"		
axisIV_1	DSM Axis IV category: problems related to legal system/crime		{Uncertain, Missing}
axisIV_2	DSM Axis IV category: problems with primary support group		

Name	Description	Component	Observed Values
axisIV_3	DSM Axis IV category: economic problems		
axisIV_4	DSM Axis IV category: occupational problems		{Yes, No, Missing}
axisIV_5	DSM Axis IV category: other psych/social/environmental problems		
axisIV_6	DSM Axis IV category: problems with access to health care services		
axisIV_7	DSM Axis IV category: educational problems		
axisIV_8	DSM Axis IV category: housing problems		
axisIV_9	DSM Axis IV category: problems related to social environment		
icd9_290	ICD-9 code in the 290 family (Dementias)		
icd9_291	ICD-9 code in the 291 family (Alcoholic psychoses)		
icd9_292	ICD-9 code in the 292 family (Drug psychoses)		
icd9_293	ICD-9 code in the 293 family (Transient organic psychotic conditions)		
icd9_294	ICD-9 code in the 294 family (Other organic psychotic conditions(chronic))		
icd9_295	ICD-9 code in the 295 family (Schizophrenic disorders)		
icd9_296	ICD-9 code in the 296 family (Episodic mood disorders)		
icd9_298	ICD-9 code in the 298 family (Other nonorganic psychoses)		
icd9_299	ICD-9 code in the 299 family (Psychoses with origin specific to childhood)		
icd9_300	ICD-9 code in the 300 family (Neurotic disorders)		
icd9_302	ICD-9 code in the 302 family (Sexual deviations and disorders)		
icd9_303	ICD-9 code in the 303 family (Alcohol dependence syndrome)		
icd9_304	ICD-9 code in the 304 family (Drug dependence)		
icd9_305	ICD-9 code in the 305 family (Nondependent abuse of drugs)		
icd9_307	ICD-9 code in the 307 family (Special symptoms or syndromess, not elsewhere classified)		

Name	Description	Component	Observed Values
icd9_308	ICD-9 code in the 308 family (Acute reaction to stress)		
icd9_309	ICD-9 code in the 309 family (Adjustment reaction)		
icd9_312	ICD-9 code in the 312 family (Disturbance of conduct)		
icd9_314	ICD-9 code in the 314 family (Hyperkinetic syndrome of childhood)		
icd9_315	ICD-9 code in the 315 family (Specific delays in development)		
Q102	"Is referral/treatment needed?"		
Q115	"Does the patient have suicidal thoughts?"		
Q119	"Is this visit for a one time consultation only?"		
Q24	"Does the patient use any caffeinated products?"		
Q37	"Does the patient have delusions?"		
Q39	"Does the patient feel safe in current living situation?"		
Q47	"Is the patient currently employed?"		
Q71	"Is the patient being referred for further medical or neurological assessments?"		
Q77	"Was the mental status exam performed?"		
Q93	"Is the patient on any kind of treatment for pain?"		
Q_HALLUCIN	Composite feature; combination of 2 questions assessing whether the patient hallucinates and/or takes hallucinogens		{Yes, No, Uncertain, Missing}
Q123	"Does the patient have a history of violent behavior?"		
Q17	"Has patient ever had a period of time when he/she felt 'up' or 'high' without the use of substances?" (Bipolar assessment)		
Q18	"Has patient ever had periods of being persistently irritable for several days or had verbal/ physical fights that seemed clearly out of character?" (Bipolar assessment)		
Q22	"Has it been more than 6 months since the loss of a loved one and does grief continue to significantly interfere with the patients daily living?" (Complicated grief assessment)		

Name	Description	Component	Observed Values
Q29	"Does the patient use cocaine?"		
Q32	"Does the patient have trouble learning new information?" (dementia assessment)		
Q33	"Has anyone told the patient they are concerned the patient has memory problems?" (dementia assessment)		
Q34	"Has the patient had periods of time lasting two weeks or longer in which they felt little nterest or pleasure in doing things or they had to push themselves to do things?" (depression assessment)		
Q35	"Has the patient had periods of time lasting two weeks or longer in which they felt sad, down, or depressed?" (depression assessment)		
Q38	"Did the patient endorse thoughts of harm to self or others during today's session?"		
Q40	"Does the patient have chronic high risk?"		
Q43	"Does the patient think they have an eating disorder?" (eating disorder assessment)		
Q44	"Has the patient had periods of time during which they were concerned about eating or their weight?" (eating disorder assessment)		
Q49	"Is there a family history of substance abuse?"		
Q51	"Is the patient under financial stress?"		
Q52	"Has the patient had times when they worried excessively about day to day matters for most of the day more days than not?" (GAD assessment)		
Q53	"Does the patient gamble?"		
Q58	"Does the patient have a history of drug use?"		
Q63	"Does the patient have a history of brain injury?"		
Q64	"Does the patient have a history of military service?"		
Q65	"Does the patient have a history of non- suicidal, self-injurious behavior?"		
Q66	"Does the patient have a history of outpatient psychiatric treatment?"		

Name	Description	Component	Observed Values
Q70	"Is the patient at risk of losing their current housing?"		
Q73	"Does the patient have any learning disabilities?"		
Q74	"Does the patient have a history of legal problems?"		
Q75	"Does the patient use marijuana?"		
Q80	"Does the patient have other repetitive unwanted thoughts or behaviors that are non-functional and difficult to stop (e.g. excessive preoccupation with appearance, motor or vocal tics)?" (OCD assessment)		
Q81	"Does the patient struggle with repetitive unwanted thoughts or behaviors for at least one hour per day?" (OCD assessment)		
Q82	"Does the patient use opiates?"		
Q84	"Does the patient use any other substances?"		
Q85	"Has the patient had episodes of sudden intense anxiety with physical sensations such as heart palpitations, trouble breathing, or dizziness that reached a peak very quickly and presented without warning?" (panic assessment)		
Q87	"Does the patient often have thoughts that make sense to them but that other people say are strange?" (psychosis assessment)		
Q88	"Has the patient has unusual experiences that are hard to explain?" (psychosis assessment)		
Q89	"Does the patient experience trauma related flashbacks or recurrent dreams/ nightmares?" (PTSD assessment)		
Q90	"Does the patient feel themselves getting very upset whenever they are reminded of their traumatic experience?" (PTSD assessment)		
Q97	"Does the patient use any prescription medications for non-medical purposes?"		
Q0	"Does the patient have longstanding problems sustaining their attention in activities that are of mediocre		

Name	Description	Component	Observed Values
	interest to them?" (ADHD assessment)		
Q1	"Does the patient have persistent fear triggered by specific objects (phobias) or situations (social anxiety) or by thought of having a panic attack?" (anxiety disorder assessment)		
Q101	"Does the patient have a history of inpatient psychiatric treatment?"		
Q106	"Is the patient on any sedative-hypnotics?"		
Q113	"Is the patient on any stimulants?"		
Q114	"Does the patient have a history of suicidal behavior?"		

Appendix 2. Final Ordinal Multiple Bayesian Network (OMBN) and modeling process diagram

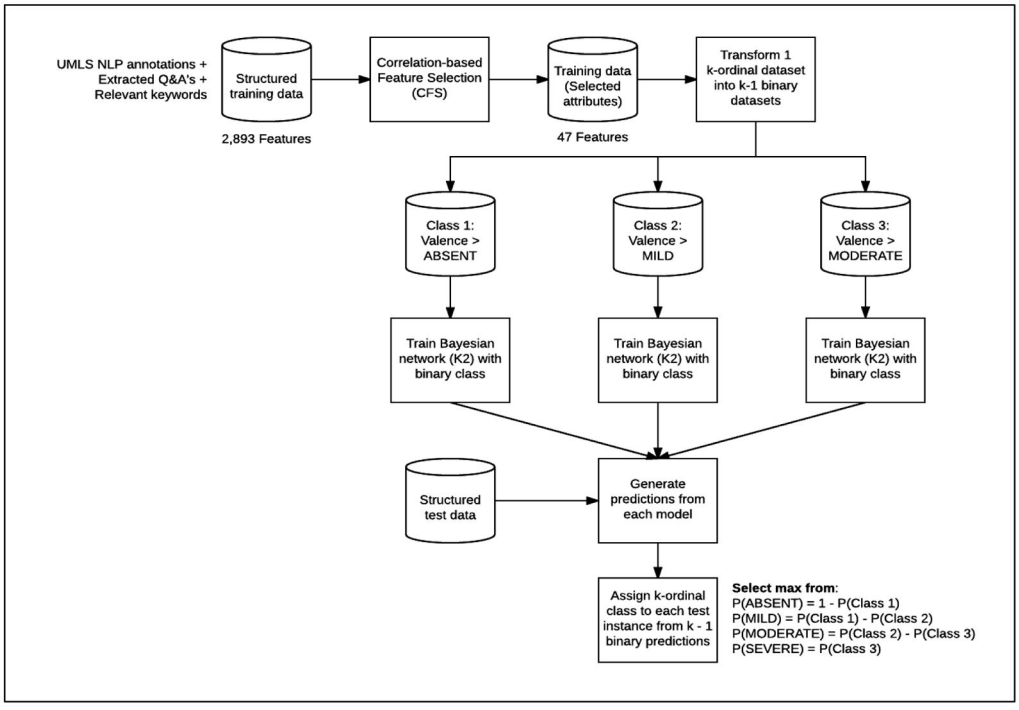


Figure 1.
OMBN modeling process diagram

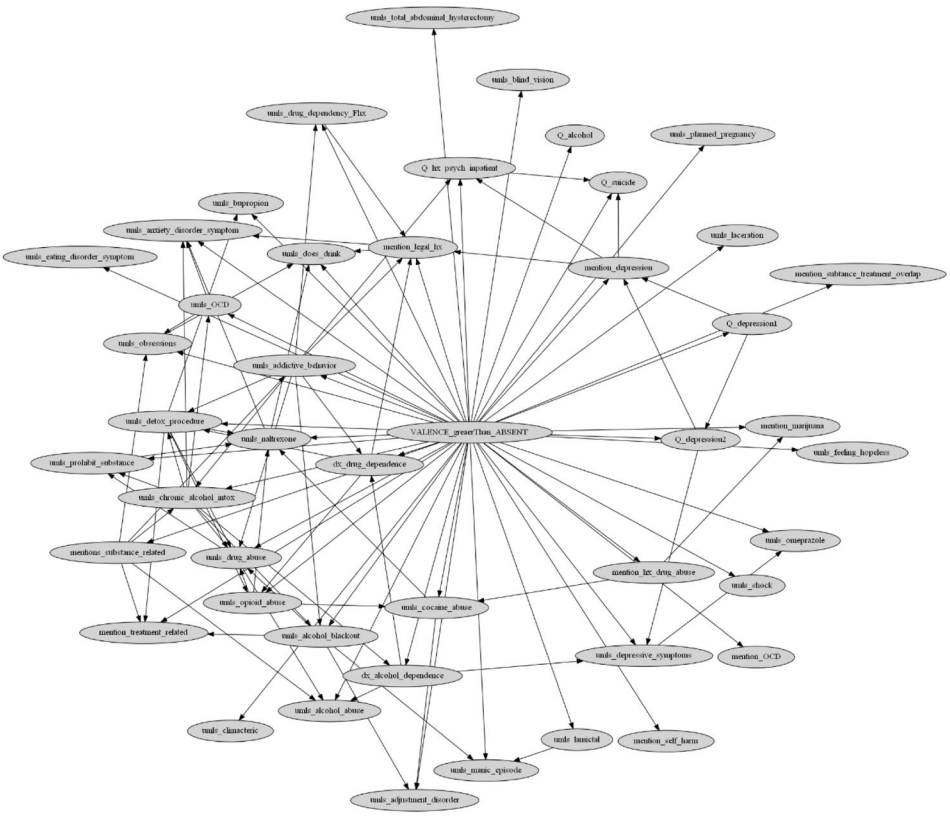


Figure 2.

A Bayesian network as a part of the Ordinal Multiple Bayesian Network (OMBN) constructed based on K2 algorithm for classifying a report to Absent or Non-Absent class.

Figure 3.

A Bayesian network as a part of the Ordinal Multiple Bayesian Network (OMBN) constructed based on K2 algorithm for classifying a report to Mild or Great-than-Mild class.

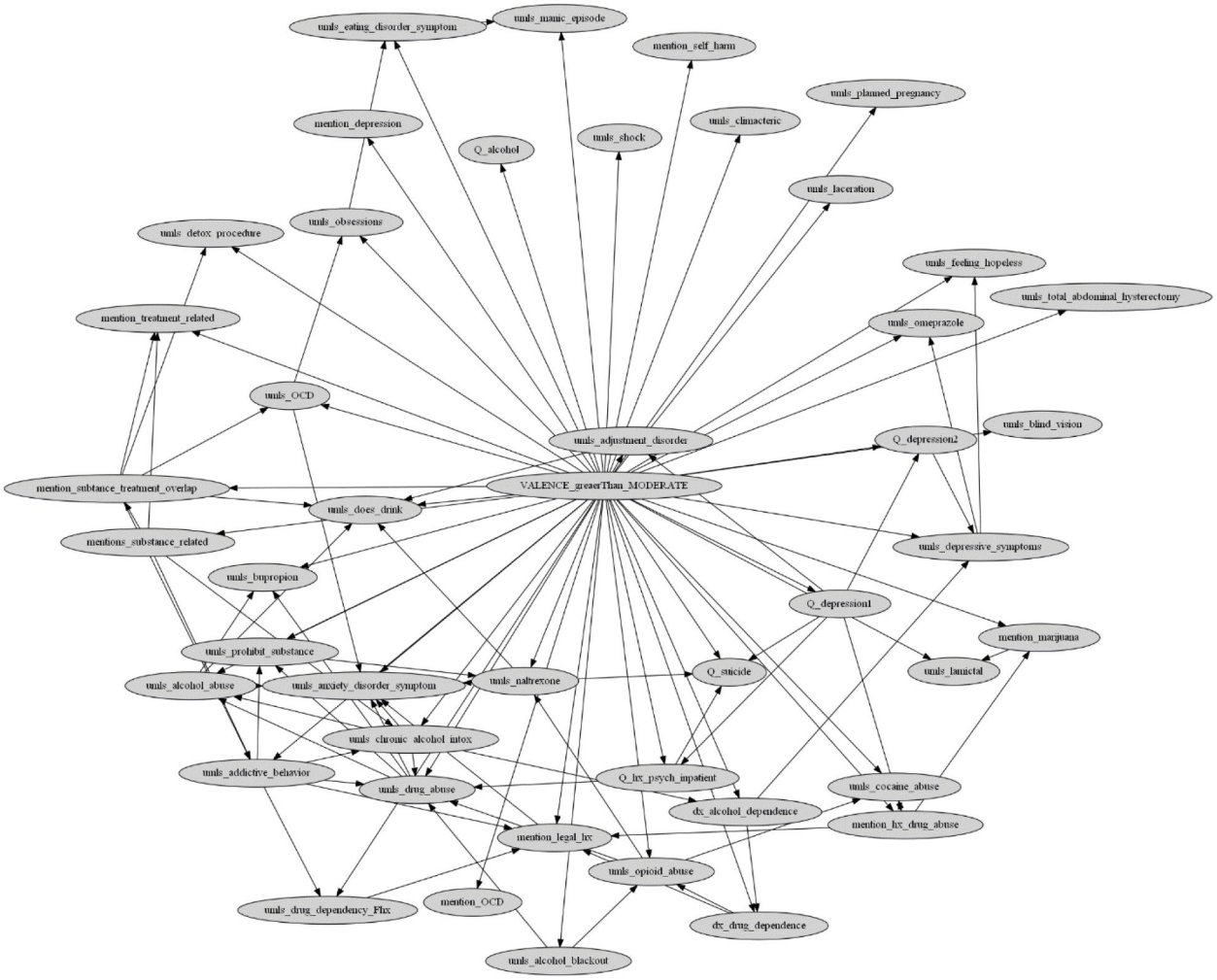


Figure 4.

A Bayesian network as a part of the Ordinal Multiple Bayesian Network (OMBN) constructed based on K2 algorithm for classifying a report to Moderate or Severe class.

Article Highlights

- Proposed a method to automatically classify symptom severity in psychiatric reports
- Question-answers from reports are the most important source of information
- Best predictive models automatically selected features prevalent in literature

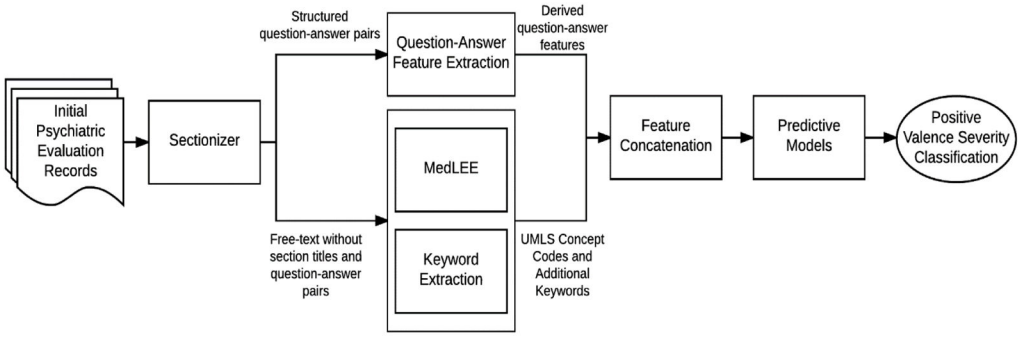


Figure 1.
Framework for classifying initial psychiatric evaluation records

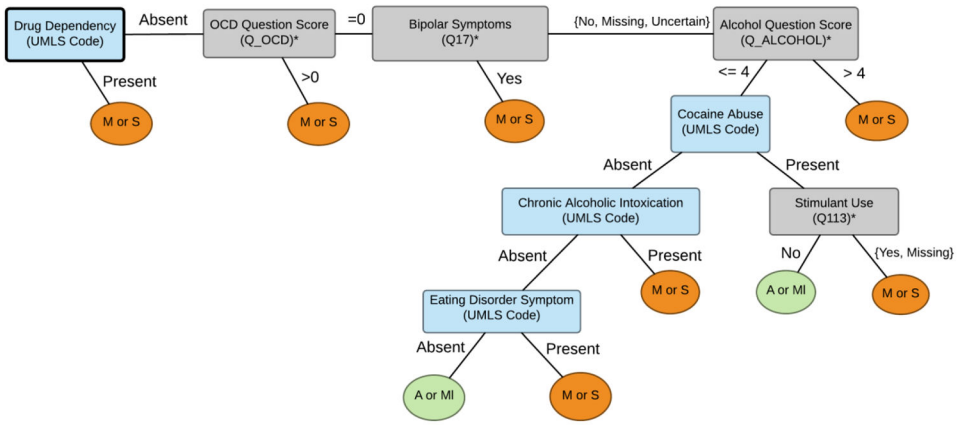


Figure 2.

Decision tree for Absent(A) or Mild(Ml) vs. Moderate(M) or Severe(S).

***** Complete description of feature available in Appendix 1.

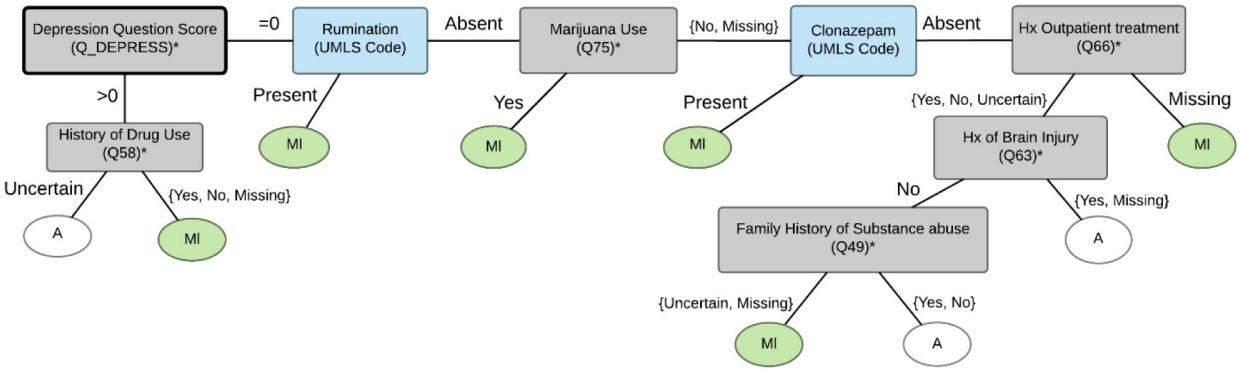


Figure 3.

Decision tree for Absent (A) vs. Mild (Ml).

***** Complete description of feature available in Appendix 1.

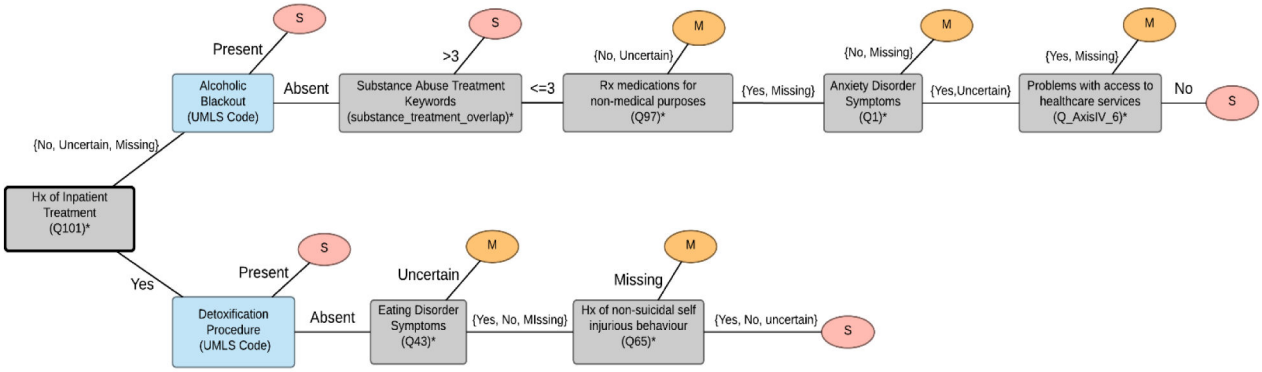


Figure 4.

Decision tree for Moderate (M) vs. Severe (S).

***** Complete description of feature available in Appendix 1.

Table 1

Final scores for predictive models

	***INMAE* (%)**									
	Training Set					**Final Test Set**				
Class	***DT SD- MC***	***DT Ab-MR***	***OMBN***	***HMBN***	***Linear SVM***	***DT SD-MC***	***DT Ab-MR***	***OMBN***	***HMBN***	**Linear SVM**
Absent	91.67	91.67	91.67	87.5	60.42	88.17	86.02	86.02	84.95	86.02
Mild	86.11	84.72	90.28	91.67	87.5	79.07	79.65	84.88	80.23	83.14
Moderate	76.79	83.93	69.64	75	76.79	79.35	75	69.56	68.48	66.3
Severe	76.19	64.29	69.05	80.95	59.52	83.65	88.05	81.76	89.31	78.62
Total	**82.69**	81.15	80.16	83.78	71.06	**82.56**	82.18	80.56	80.74	78.52

Table 2

Performance impact of different features for best performing algorithm (DT SD-MC) on the test set. Q&A: Question-answer pairs

Class	Q&A	MedLEE	Keywords	Q&A + MedLEE	Q&A + Keywords	MedLEE + Keywords
Absent	**90.32**	84.95	56.99	87.1	84.95	74.19
Mild	77.91	77.9	**85.47**	77.91	78.49	76.16
Moderate	72.83	63.04	68.48	**75**	65.22	61.96
Severe	76.1	77.99	71.07	**88.68**	73.58	72.96
Total	79.29	75.97	70.5	**82.17**	75.56	71.32