Post-Traumatic Stress Disorder (PTSD) Ontology and Use Case

Bryan Travis Gamble[a,b*], Matthew Brush[a],
Aaron Cohen[a], Melissa Haendel[a]
[a]Oregon Health & Science University, Portland, OR, USA
gamble@ohsu.edu[*]

Thomas Rindflesch[e]
[e]National Library of Medicine,
National Institutes of Health, Bethesda, MD, USA

Maryan Zirkle[b], Dezon Finch[c], Ruth Reeves[d],
David Hickham[b], Stephen L Luther[c]
[b]Portland VA Healthcare System, Portland, OR,USA
[c]James A Haley Veterans Hospital, Tampa, FL, USA
[d]Tennessee Valley VA Healthcare System,
Nashville, Tennessee, USA

Samah Jamal Fodeh[f,g*], Jonathan Bates[f],
Cynthia Brandt[f,g], Kei-Hoi Cheung[f,g*]
[f]VA Connecticut Healthcare System, West Haven, CT, USA
[g]Yale School of Medicine, Yale University,
New Haven, CT, USA
samah.fodeh@yale.edu[*], kei.cheung@yale.edu[*]

***Abstract*—Ontologies play an increasingly important role in annotation, integration, and analysis of biomedical data. In this paper, we describe the design and development of a Post-Traumatic Stress Disorder (PTSD) Ontology and how we can use this ontology as a controlled vocabulary for supporting automatic annotation of clinical text. The automated annotation is performed using a natural language processing (NLP) tool called YTEX. In addition, we demonstrate how we can use the concepts and relationships defined in the PTSD Ontology to perform data summarization and categorization.**

Keywords— PTSD, mental disorder, natural language processing, data categorization, clinical note analysis

I. INTRODUCTION

Ontology development is motivated by providing semantic context, automated reasoning and annotation, data mining and analysis, and decision-making support. In addition to ontological efforts at the Unified Medical Language System (UMLS; http://www.nlm.nih.gov/research/umls), the National Center for Biomedical Ontologies (NCBO) has developed a repository called "BioPortal" [1] that allows both manual and programmatic access to several hundreds of biomedical ontologies including some of those from the UMLS. To promote quality and standard practice, the Open Biological and Biomedical Ontologies (OBO) Foundry [2] has established a set of principles for ontology development with the goal of creating a suite of orthogonal interoperable reference ontologies in the biomedical domain. While the development of ontologies is growing and maturing, there is still a need for expanding existing ontologies or developing new interoperable ontologies that describe new domains of knowledge in the biomedical domain. In addition, the utility of ontologies in clinical or health information applications has not yet been fully demonstrated.

In this paper, we describe the development of a new ontology in the domain of Post-Traumatic Stress Disorder (PTSD). The American Psychiatric Association defines PTSD as a condition occurring from exposure to a trauma that impacts the physical integrity or life of the individual or of another person [3]. It is considered normal for an individual to have a strong reaction to a traumatic event but the effects should decrease over time when the threat is no longer present. However, people with PTSD continue to experience extreme reactions and symptoms even after the trauma is no longer present [4]. According to the National Center for PTSD, 7-8% of the population in the U.S. will have a form of this disorder at some point in their lives [5].

Much of the significant clinical details of a health condition is usually recorded in unstructured clinical notes as part of the electronic health records (EHR). This clinically useful information is typically abstracted using natural language processing (NLP) and machine learning techniques. However, because these automated methods are blind to the sublanguages used to describe the various health conditions in the notes, the knowledge coded in domain-specific ontologies can serve as a useful guide to the automated process of data abstraction. As similar consortiums have driven toward standardizing the representation of phenotype data in other domains, such as efforts with eMERGE (electronic MEdical Records and GEnomics:https://www.mc.vanderbilt.edu/victr/dcc/projects/acc/index.php/Main_Page) [6], we are putting forth similar interventions to improve the limited existing coverage of PTSD.

In this paper, we provide a use case showing how the PTSD Ontology can be used to support automatic annotation of clinical text. In addition, we discuss how the PTSD concepts and relationships can be used to perform data categorization.

II. ONTOLOGY DEVELOPMENT

A. PTSD Ontology

Knowledge representation of PTSD is limited by heterogeneous yet overlapping terminologies and the subjective narrative patient information in electronic clinical notes. The lack of contextual analysis with this unstructured evidence is a barrier to the understanding of PTSD symptoms

* Co-first-authors who contributed equally. Correspondence should be addressed to Bryan Gamble, Samah Jamal Fodeh, and Kei-Hoi Cheung

and the evaluation of treatment effectiveness. To surmount these hurdles, we are developing an ontology to capture knowledge relevant to PTSD symptoms and treatments. The PTSD Ontology is being built to share domain knowledge of relevant concepts in a formal framework representation and capture the semantic relationships between those concepts. For instance, the semantic relation isA represents subclasses of specific PTSD symptom clusters defined within the framework. The ontology is being developed to specify the concepts, relationships, instances, and axioms explicitly to enable more precise search and reasoning about this data. Such an ontological framework allows domain knowledge to be shared and reused across applications.

B. Design elements and principles

This ontology is developed in the Web Ontology Language (OWL) using Protégé Version 4.3. Our approach plans to make use of existing knowledge bases and ontologies via an ontological import. Currently, the PTSD Ontology includes imports from existing data collections such as Systematized Nomenclature of Medicine - Clinical Terms (SNOMED-CT: http://purl.bioontology.org/ontology/SNOMEDCT), Symptom Ontology (SYMP:http://purl.bioontology.org/ontology/SYMP), Ontology of General Medical Science (OGMS: http://purl.bioontology.org/ontology/OGMS), National Cancer Institute (NCIT: http://purl.bioontology.org/ontology/NCIT), Medical Dictionary for Regulatory Activities (MEDDRA: http://purl.bioontology.org/ontology/MEDDRA), and the upper-level Basic Formal Ontology (BFO: http://purl.bioontology.org/ontology/BFO) among other resources [1].

Our design of the PTSD Ontology follows the principles of the OBO Foundry [2] in order to ensure interoperability with the existing reference ontologies. The PTSD Ontology shares many concepts that exist in other ontologies and reference terminologies. For example, the PTSD Ontology contains concepts defined in the Symptoms Ontology but with added context of traumatic exposures that accompany both behavioral and physiological symptomatology spectrum. An application that was supported with Symptoms Ontology as a reference terminology would be interoperable with similar but modified concepts defined in the PTSD Ontology. The goal is for other researchers to be able to implement relevant concepts and relationships in order to systematically share, reuse, and alleviate inconsistencies in disparate data sets across the PTSD community. Controlled vocabulary resources, literature reviews and expert panels form the building blocks of our ontological foundation. The existing coverage was excellent for building a terminology base but was limited and inadequate in completeness to meet the needs for our current and future use case implementations. Lastly, for increased coverage, the PTSD Ontology incorporated annotations with symptom and treatment terms extracted from mental health notes of patients with PTSD extracted from the Veterans' Health Administration Corporate Data Warehouse (VHA CDW). Our ontological contents (including concepts, synonyms, relationships and their hierarchical organization) has been validated by clinicians and PTSD domain experts.

C. Ontological structure and content

The current in progress version of the PTSD Ontology is available for public download at http://code.google.com/p/ptsd-ontology/. Coverage of the ontology is currently directed to variations in symptoms and treatments. While the work is on-going, the purpose of the conceptual PTSD model design is to support: 1) retrieval, collection, and sharing of information; 2) natural language processing (NLP) tasks; and 3) ontology-driven information extraction (IE) for automated accumulation of symptoms and treatments located within the narrative portion of a patient's EHR encounter data. The ontology is being designed to account for a wide range of treatments and to recognize the specificity and intensity of symptoms. Currently, the PTSD Ontology consists of 219 symptom classes and 367 treatment classes. PTSD symptoms are arranged in clusters according to definitions in the Diagnostic and Statistical Manual of Mental Disorders, 5th ed. [3]. Clusters include stressors, intrusion symptoms, avoidance and numbing, negative alterations in cognitions and mood, alterations in arousal and reactivity, functional significance and dissociative symptoms. A subset of avoidance subclasses is displayed in **Figure 1**. It is important to semantically distinguish these variations in symptoms as they translate directly into the diagnosis of disease and type and breadth of clinical care. While the variations in symptoms are applicable to multiple cohorts, the context of this framework is derived from adult patients with traumatic stress reaction treated in a Veterans Healthcare Administration (VHA) clinical setting. This symptom grouping establishes parameters necessary for the semantic understanding of assessment, diagnosis, and management of symptoms. Similarly, concepts describing treatment interventions are arranged in categories designated in the Veteran Affairs/Department of Defense (VA/DoD) PTSD evidence-based practice management guidelines [8].

Five primary therapeutic categories of pharmacotherapy, psychological, psycho educational, psychosocial, and case management are shown in **Figure 2**. Knowledge about the variations in available prescribed treatments for PTSD can further enhance our ability to comparatively evaluate their relationships and effectiveness on treating the symptoms of this illness. This organization provides structure for symptom-

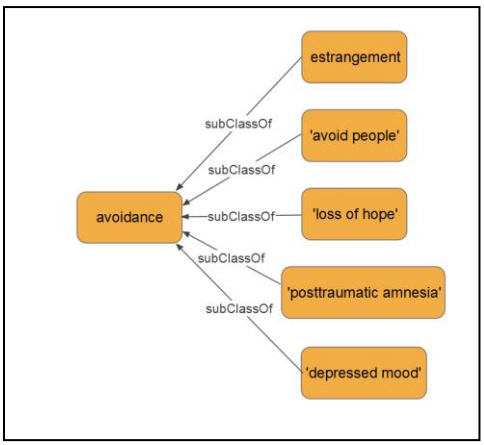

Fig.1Subset of avoidance subclasses.

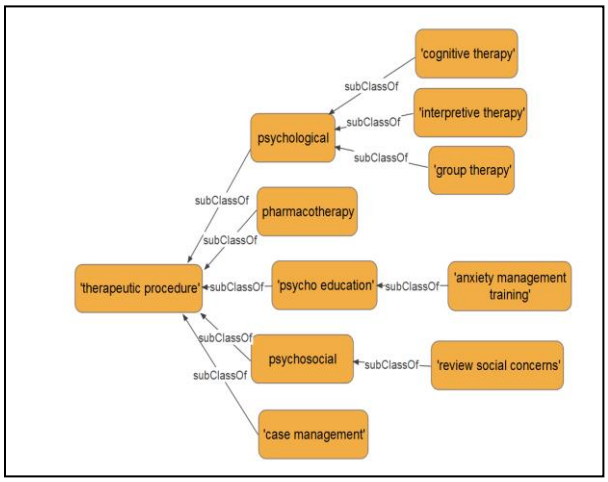

Fig. 2. Subset of primary therapeutic categories classes

specific management supporting precision of information retrieval and classification. The ontology can be customized to support personalized therapeutic approaches when treating heterogeneous symptoms that persist at the individual patient-level. The hierarchical arrangement of symptoms and treatments allows representation of data using parent/child relationships and fosters organization of information to leverage automated retrieval. Subclass relations establish hierarchical relationships between classes, while other properties are used to classify data along other axes. **Figure 3** shows some of the classes and a high-level overview of treatment classes establishing the "treats" property displaying this non-hierarchical relationship with specific symptoms described in the ontology. As the gaps in current understanding of the disorder are addressed, it is important for our structure to set parameters that foster contextual collaboration. The framework of the PTSD Ontology aids establishing a consensus on the semantic understanding of terms and relationships used to describe this disorder.

III. USE CASE

In this section, we discuss two applications of the PTSD Ontology: automatic annotation of PTSD clinical notes and categorization of notes' contents based on the hierarchical relationships defined in the ontology. A subset of the PTSD Ontology was obtained by loading the ontology into Protégé

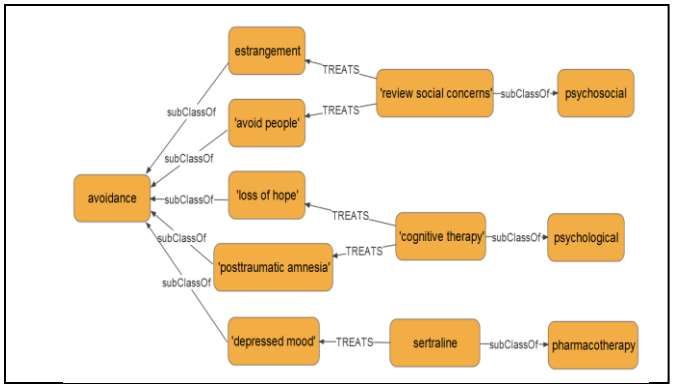

Fig. 3. Established relationships between classes.

and retrieving the terms and relationships we wanted using SPARQL queries. The query output was produced in the tab-delimited format.

A. NLP annotation use case

Projects like Annotator (http://bioportal.bioontology.org/annotator) and ODIE (http://bioontology.stanford.edu/ODIE-project) enable the use of biomedical ontologies in natural language processing (NLP). In the context of clinical NLP, YTEX [9] was used to automatically annotate clinical notes involving different medical conditions (e.g., fall and lung cancer). YTEX is an extension to the clinical Text Analysis and Knowledge Extraction System (cTAKES) to derive robust feature sets from NLP pipelines [10]. The output components generated by YTEX include words, concepts, phrases, sentences and annotations of concepts, which can be stored in a relational database. Among these components, only concepts were extracted and used as features. The dictionary component of YTEX is composed of UMLS clinical concepts; it feeds into the named entity recognition module to annotate concepts. This dictionary can be customized and its contents can be replaced by the vocabulary of the user's interest. Since the goal of this study is to identify only PTSD treatment and symptom related concepts and because the UMLS includes non-PTSD concepts from a variety of sources, we replaced the built-in dictionary of YTEX with a pared-down set of concepts obtained from the PTSD Ontology.

An excerpt from an actual outpatient progress note for a patient with PTSD is shown in the YTEX annotation viewer in **Figure 4**. Spans of text that were mapped to concepts (symptoms or treatments) in the PTSD Ontology vocabulary are highlighted by the annotation viewer.

B. Data analysis

The developed controlled vocabulary for PTSD will serve as valid components for data categorization of PTSD treatments. As described in Section III.A, the vocabulary encoded in the newly established PTSD Ontology was utilized to detect treatments of PTSD mentioned in clinical notes. The treatment concepts detected and extracted for each clinical note were used to compose a bag of concepts (BOC) representation of the notes. In this representation the concepts are arranged in a matrix where the rows are the clinical notes and the columns are the treatment concepts. This representation can be effectively utilized in subsequent machine learning and data analysis tasks. In this work, we demonstrate the utility of the PTSD Ontology by building a condensed representation of clinical notes using the ontology's hierarchical relationships.

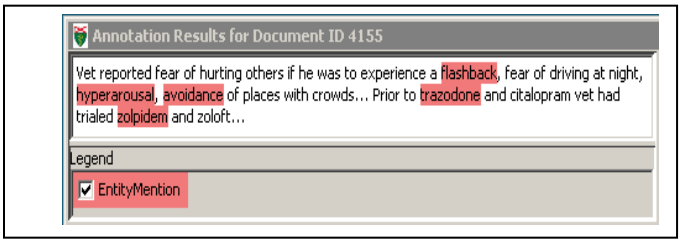

Fig. 4. Concept annotations generated by YTEX.

The advantage of the new data representation is a reduction in dimensionality i.e. the size of the feature set (concepts) to reduce complexity in the analysis of large volumes of notes. Because the BOC extracted from YTEX output (see Section III.A) contains a variety of concepts, some of which are relevant to our interests and others not, we use the hierarchical structure of the PTSD Ontology to help remove the irrelevant concepts. We present a process to transform the concepts extracted from the notes into a more general, less granular set of concepts by integrating knowledge from the PTSD Ontology. This transforms the representation of the text notes from concepts to more abstract categories. The benefit of this transformation is three-fold: First, it reduces the complexity and sparsity of data analysis by decreasing the dimensionality of the space. Second, it provides a focused/targeted analysis of the notes by removing the features that do not belong to the categories of our interest and not relevant to the clinical use-case which could obscure the analysis. Third, it may reveal new categories to capture and conceptualize the data for better understanding. We transformed the BOC representation of the PTSD clinical notes to the Bag Of Categories (BOCat) representation, where the categories are the types of PTSD treatments. A significant reduction of dimensionality is achieved using BOCat. In the BOC representation, there were 367 concepts to describe the notes whereas in the BOCat representation, the feature space is compressed into a higher ontological level consisting of 6 treatment categories (dimensions) and the notes are described using these higher level PTSD treatment concepts in the ontology. To discard irrelevant concepts i.e. symptoms from the BOC and generate the focused BOCat of treatment only, each concept is mapped to its treatment category using the hierarchical relationships in the PTSD Ontology. This forms a filter, wherein concepts that do not belong to a treatment type are dropped from the analysis. In addition to noise and dimensionality reduction, the BOCat representation assigned weights to each category of treatments in the notes. The weight of a treatment type is calculated for a particular clinical note by summing the frequencies of all concepts belonging to that type of treatment. This information, typically documented exclusively in the narrative text, indicates how often a treatment type is documented in a clinical note and how effective it might be for a patient's existing PTSD symptoms.

IV. DISCUSSION

The data used in this analysis consisted of complete sentences of narrative text from clinical documentation that included much "noise" such as abbreviations, misspellings, negations. An important goal of this project is to make explicit the assumptions in PTSD clinical note data and thereby reduce the ambiguity in concepts that describe symptoms and treatments in this domain. In mental health, and more specifically in anxiety disorders, concepts are often shared with slight modifications corresponding to different contexts. For example, many PTSD re-experiencing symptoms are similar to obsessive-compulsive disorder (OCD) symptoms but with the differential details relating the recollections to the traumatic exposure. As we continue to develop more complex use cases, we are aware of the impedance mismatch between ontologies and information models. As described by Ceusters, "terms in ontologies refer to universals, where clinical histories consist overwhelmingly of representational units that refer to instances [11]." The subjective nature of symptom identification can be problematic for ontology development within the PTSD domain inasmuch as the same symptoms may be associated with disparate formal diagnoses and treatment recommendations. The development of description logic within ontology, customized to the domain can help overcome another important obstacle to data interoperability in mental health research: the use of different assessments and scales for measuring symptoms and assisting in diagnosis [12].

Validation, feature reduction, and identification of PTSD clinical data categorization are currently underway. Future analysis will compare ontology coverage and accuracy with existing terminologies including the UMLS. Continued development will assist in analyses of clinical relationships between symptoms and treatments. The PTSD Ontology can potentially facilitate research collaborations on varied assessments and structured interviews. Ontological representation and reasoning may help improve prediction, prognosis, and understanding of this complex disorder.

ACKNOWLEDGMENT

We gratefully acknowledge support from Oregon Health & Science University, the Department of Veteran Affairs Healthcare System, and Yale University School of Medicine. This study was supported in part by the Intramural Research Program of the National Institutes of Health, National Library of Medicine

REFERENCES

[1] Noy,N.F. et al. (2009) BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic acids research. 2, W170-173.

[2] Smith,B et al. (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nature biotechnology. 11, 1251-1255.

[3] American Psychiatric Association. (2013) Diagnostic & Statistical Manual of Mental Disorders,5th ed. American Psychiatric Association, Washington, DC.

[4] Norris,F. and Tracy,M. and Galea,S. (2009) Looking for resilience: Understanding the longitudinal trajectories of responses to stress. Social Science & Medicine, 03(043).

[5] U.S. Department of Veterans Affairs. How Common is PTSD? 30 Jan. 2014. Web. 01 Jul. 2014. http://www.ptsd.va.gov/public/PTSD-overview/basics/how-common-is-ptsd.asp).

[6] Pathak,J et al. (2011) Mapping clinical phenotype data elements to standardized metadata repositories and controlled terminologies: the eMERGE Network experience. J Am Med Inform Assoc.18(4):376-386.

[7] McGuinness,D.L. and Van Harmelen,F. (2004) OWL web ontology language overview. W3C recommendation, 10, 2004-03.

[8] Department of Veterans Affairs. (2004) VA/DoD clinical practice guideline for post-traumatic stress management. Department of Defense. 2010 update.

[9] Garla,V. et al. (2011) The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc. 18(5):614-20.

[10] Savova.G.K. et al. (2010) Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 17(5):507-13.

[11] Ceusters,W. and Smith,B. (2010) Foundations for a realist ontology of mental disease. J. Biomedical Semantics, 1, 10.

[12] Luther,S. et al. (2011) Using statistical text mining to supplement the development of an ontology. J Biomed Info. 44, S86-S93.