Journal of Biomedical Informatics 44 (2011) S86–S93

Contents lists available at SciVerse ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Using statistical text mining to supplement the development of an ontology

Stephen Luther [a,b,*], Donald Berndt [a,c,1], Dezon Finch [a,b], Matthew Richardson [a,b], Edward Hickling [a,b], David Hickam [a,d,e]

[a] Consortium for Healthcare Informatics Research (CHIR)
[b] VA HSR&D/RR&D Center of Excellence: Maximizing Rehabilitation Outcomes, Tampa, FL, United States
[c] College of Business, University of South Florida, Tampa, FL, United States
[d] HSR&D Research Enhancement Program, Portland VA Medical Center, Portland, OR, United States
[e] Department of Medicine, Oregon Health and Science University, Portland, OR, United States

ARTICLE INFO

Article history:
Available online 15 November 2011

Keywords:
Text mining
Controlled vocabulary
Ontology development
Post-traumatic stress disorder (PTSD)

ABSTRACT

Statistical text mining was used to supplement efforts to develop a clinical vocabulary for post-traumatic stress disorder (PTSD) in the VA. A set of outpatient progress notes was collected for a cohort of 405 unique veterans with PTSD and a comparison group of 392 with other psychological conditions at one VA hospital. Two methods were employed: (1) "multi-model term scoring" used stepwise logistic regression to develop 21 separate models by varying three frequency weight and seven term weight options and (2) "iterative term refinement" which used a standard stop list followed by clinical review to eliminate non-clinical terms and terms not related to PTSD. Combined results of the two methods were reviewed by two clinicians resulting in 226 unique PTSD related terms. Results of the statistical text mining methods were compared with ongoing efforts to identify terms based on literature review, focus groups with clinicians treating PTSD and review of an existing vocabulary, lending support to the contributions of the STM analyses.

Published by Elsevier Inc.

1. Introduction

Post-traumatic stress disorder (PTSD) is a common clinical problem in the Veterans Health Administration (VA), particularly among men and women who have served in Operation Enduring Freedom (Afghanistan) or Operation Iraqi Freedom (OEF/OIF). It is estimated that the prevalence of PTSD among OEF/OIF veterans may be as high as 20% [1]. PTSD is generally a lifetime disorder, and its clinical manifestations are diverse. A primary goal of clinical management is relief of symptoms, and the success of treatment methods is measured by changes in symptoms and functioning over time. In the routine follow-up of PTSD patients, clinicians annotate the presence and severity of symptoms in progress notes that provide a record of their clinical care. Thus, the VA's electronic health record (EHR) provides detailed information about the clinical status of patients who are followed for PTSD in the VA system. However, most of this information (which is contained in narrative text) is not accessible through administrative data sources that are easily searched and extracted for analysis. Because there is substantial variability in successful alleviation of the symptoms of PTSD, the lack of good longitudinal data has hampered clinical efforts to improve care. Better methods to capture the clinical information in VA progress notes promises to meet this important need.

The VA Consortium for Health Informatics Research (CHIR) is organized as a multi-disciplinary group of collaborating investigators located at VA sites distributed across the United States. The primary participating VA sites include Portland, Palo Alto, Salt Lake City, Indianapolis, Nashville, Tampa, West Haven, and Boston. The academic institutions affiliated with each of these VA facilities serve as research partners. Disciplines and concentration areas represented by CHIR investigators include knowledge representation, natural language processing, machine learning, biostatistics, clinical epidemiology, applied informatics, and health services research.

The CHIR is conducting two multi-year, applied studies which address clinical domains of high priority for veterans, methicillin-resistant Staphylococcus aureus (MRSA) infection and post-traumatic stress disorder (PTSD). These major projects are designed to drive advancement of natural language processing (NLP) methods and to lead to clinical applications to improve the quality of care. The current study is part of the PTSD project. The PTSD project will measure the potential of unstructured text to provide information about the clinical course and symptom

* Corresponding author. Address: Consortium for Health Informatics Research (CHIR) and the VA Center of Excellence: Maximizing Rehabilitation Outcomes, Tampa, FL, United States.
E-mail addresses: steve.luther@med.va.gov (S. Luther), dberndt@usf.edu (D. Berndt), dezon.finch@va.gov (D. Finch), matthew.richardson@va.gov (M. Richardson), edward.hickling@va.gov (E. Hickling), david.hickham@va.gov (D. Hickam).
[1] Fax: +1 813 558 7616.

1532-0464/$ - see front matter Published by Elsevier Inc.
doi:10.1016/j.jbi.2011.11.001

variation among veterans who receive clinical care from the VA for PTSD. The lack of adequate codified data on symptoms of PTSD and on psychosocial correlates of PTSD has been a barrier to evaluating the effectiveness of clinical strategies for management of this pervasive condition.

The first goal of the CHIR PTSD project is to define the vocabulary used by clinicians to describe the clinical course of veterans with PTSD to provide a framework upon which NLP-based concept extraction will be built. This vocabulary will eventually be expanded into an ontology related to PTSD. The process of developing the vocabulary focused on deductive techniques and qualitative methods including literature review, focus groups and review of existing vocabulary resources. However, access to large corpora and the evolution of machine learning techniques permit using knowledge discovery techniques to supplement the deductive approach. Cimiano [2] suggests that the acquisition of domain knowledge from data (ontology learning) is made up of sequential steps that can be organized as a layer cake according to increasingly complex subtasks. The bottom layer (foundation) of the cake represents the acquisition of relevant terminology, followed by identification of synonym terms, formation of concepts, hierarchical organization of concepts, learning relationships among terms, hierarchical organization of the relationships, instantiation of axiom schemata and definition of arbitrary axioms. Here we describe efforts to use machine learning techniques to identify terms, the task at the bottom layer of Cimiano's cake. We report results from the use of statistical text mining (STM) to extract terms related to post-traumatic stress disorder (PTSD) from outpatient progress notes to supplement more traditional deductive steps being taken to develop the vocabulary.

We employ two STM techniques to extract terms for clinical text. The goal is to use a statistical approach, largely free of human biases, to provide an orthogonal means of term discovery to generate a supplemental term list for clinician review. The STM or "bag-of-words" approach does not rely on any existing controlled vocabularies, so the term discovery is largely unconstrained and could uncover surprising candidate terms. While NLP techniques, coupled with data mining could also be used for some tasks, any discoveries are limited to the vocabulary being used for term look-ups. Since our task was to assist, rather than replace methods that already use literature review for discovery, a statistical approach seemed the most appropriate.

Perhaps the most novel aspect of this approach is that the target of STM was clinical progress notes, not the medical literature. Since many applications of the resulting PTSD vocabulary will be clinically focused, using actual notes seems likely to uncover terms based on intended usage. In fact, prior research has shown that there are many sublanguages within medicine [3]. This is probably true within the literature, as well as within the various clinical note types authored by providers across medical specialties. However, the varied language within notes provides an especially relevant source of terms for vocabulary construction and subsequent text mining applications. For this work, the corpus contained 41 different note types, representing somewhat different language patterns.

1.1. Background

Analyses of large corpora are increasingly used to enhance vocabulary and ontology development. Some investigators have used automated text extraction to identify relationships among terms in a domain and to build ontologies. For example Coulet et al. describes a process of building relationships based on key pharmacogenomic entities and a syntactic parse of more than 87 million sentences from 17 million MEDLINE abstracts to systematically extract commonly occurring relationships and to map them to a common schema [4]. In another example Baneyx et al. applied natural language processing tools to two corpora, one composed of patient discharge summaries and the other being a text book, to enrich ontology building through distributional analysis and a method based on the observation of corpus sequences in order to reveal semantic relationship [5]. More commonly however, automated efforts to help develop ontologies for text have been focused on generating terms upon which ontologies can be developed. Automated techniques have been used to leverage information in the biomedical literature in support of identifying terms related to genes and gene products [6], the discovery of abbreviations and definitions [7], the creation of a dictionary of protein names [8], and as a way to supplement manually developed controlled vocabularies [9]. Fewer studies have focused on using clinical text to develop or improve controlled vocabularies. Investigators at Columbia have shown that natural language processing procedures can be used to support or enhance vocabulary development based on information available in the electronic health record [10,11]. However, we could find no published study that employed STM methods to identify terms in support of building vocabularies and ontologies.

2. Methods and materials

Statistical text mining (STM) employs inductive or data-driven algorithms that do not rely on large controlled vocabularies. We use SAS Enterprise/Text Miner for the analysis. The SAS text mining software implements several algorithms for text mining, providing a supportive environment for file processing, text parsing, transformation, dimension reduction and document analysis. We believe that the inductive analyses of the data may discover concepts that would not have been identified by the clinical team. Two main approaches are used to surface terms for possible inclusion in controlled vocabularies or ontologies. Multi-model term scoring uses a fairly large group of different models to uncover terms, with a synthetic score ranking terms across all models. The second method, iterative term refinement, uses a single predictive model (selected after parameter tuning) to surface a set of terms. An iterative cycle of clinician review is then used to remove terms (via start/stop lists) allowing additional terms to be discovered in subsequent model building steps.

2.1. Document selection and labeling

To conduct statistical text mining, documents (progress notes) need to be labeled as containing information about PTSD or not. For this analysis, we leverage information in patient level administrative data to identify appropriate notes. A list of 5165 unique veterans of OEF/OIF who received care at one large Veterans Hospital in the southeast during FY 2007 (October 2006 to September 2007) was identified based on reports from the VA's Support Service Center Web site. From this initial pool we used multiple criteria based on information documented in administrative data stored in the EHR to identify veterans who were diagnosed with and treated for PTSD. A veteran was considered to have PTSD if he or she had a flag in her record indicating that they had been confirmed as having the condition by the VA Compensation and Benefits program (service connected), had at least two outpatient visits in the year with the primary diagnosis being listed as PTSD (ICD-9-CM code 309.81) and had PTSD listed on the problem list in the VA EHR. Based on these criteria, 405 unique veterans receiving care for PTSD during the study period were identified. We then identified a potential comparison group to be used in the analyses. Our goal was to find patients with similar psychological conditions but not PTSD. We anticipated that this would result in the identification of terms that are strongly associated with PTSD but less strongly related to other diagnoses. To do this we identified a cohort of