Paper SDA-08

Using SAS® Text Miner 4.1 to create a term list for patients with PTSD within the VA

Matthew R Richardson MSPH, MBA, James A Haley VA Hospital, Tampa, Fl
Stephen L Luther PhD, MBA, James A Haley VA Hospital, Tampa, Fl
Donald Berndt PhD, MBA University of South Florida, Tampa, Fl

ABSTRACT
SAS® Text Miner 4.1 was utilized to perform statistical text mining to supplement efforts to develop a clinical vocabulary for post-traumatic stress disorder (PTSD) in the VA. 405 veterans with PTSD were identified through administrative data within the Tampa VA and then combined with a comparison group of 392 veterans with no known PTSD symptoms or diagnosis. The patient notes from this cohort were captured to produce the dataset analyzed. Using all possible combinations of Frequency Weight and Term Weight 24 different models were run. 21 models produced usable results. In each model terms were ranked and scored based on global weights and then ranked and scored again based on the absolute value of the coefficient in a regression analysis utilizing stepwise regression. Scores for each model and each analysis were combined to conceive a master score with the highest scored term receiving the highest rank. 449 terms were identified within the regression analysis and 343 were identified in the global weight analysis for a total of 585 distinct terms. Future studies will focus on incorporating processes that identify terms that carry greater clinical relevance as well as incorporation of the improved techniques behind SAS Text Miner 4.2.

INTRODUCTION

Psychiatric disorders are among the more pervasive health outcomes resulting from wartime deployment, with Post Traumatic Stress Disorder (PTSD) being the most prevalent mental health problem reported among combat Veterans.[1] PTSD is a psychiatric condition resulting from exposure to an experience involving direct or indirect threat of serious injury or death. Symptoms include recurrent thoughts of a traumatic event, hyper alertness, anxiety and irritability.[2] While a wide variety of events can cause PTSD (e.g. floods, airplane crashes, torture, etc.) in this study, we focus on PTSD that occurs as a result of experiences in war zone deployment that is “deployment-related PTSD”. PTSD is not only common among Veterans, but sheer numbers suggest that its effects are both acute and enduring, and contribute to a considerable portion of the disability reported within the Veterans Administration (VA).

Typically, a deductive process, employing qualitative methods such as literature review, expert panels, interviews and focus groups is used to develop clinical vocabularies and ontology’s. However, given access to large corpora and the evolution of machine learning techniques there is increasing support for using knowledge discovery techniques to supplement or replace the deductive approach.[3] Here we report on some initial efforts by researchers from the Veterans Administration’s (VA) Consortium for Healthcare Informatics Research (CHIR) to use statistical text mining (bag of words) to extract terms related to post-traumatic stress disorder (PTSD). We outline a strategy that summarizes information obtained from multiple models to generate a term list for experts to review and consider.

BACKGROUND

CHIR is organized as a multi-disciplinary group of collaborating investigators located at VA sites distributed across the US. The primary participating VHA sites include Portland, Palo Alto, Salt Lake City, Indianapolis, Nashville, Tampa, West Haven, and Boston. The academic institutions affiliated with each of these VHA 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 variation among veterans who receive VHA care 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.

PTSD is a common clinical problem in the 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%.[4] 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 medical record (EMR) 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 the 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.

As part of the overall research agenda of the VA CHIR, efforts are underway to define the vocabulary used by clinicians to describe the clinical course of veterans with PTSD. A qualitative research approach is ongoing to elicit from VA clinicians information on their documentation practices and vocabulary that represents the domains of information that relate to diagnosis and treatment of veterans with PTSD. Results of this deductive process will be used to develop an ontology that will provide a framework upon which NLP-based concept extraction will be built. The purpose of this first step is to gain consensus among expert clinicians on the categories of clinical assessment relevant to the care of veterans with combat-related PTSD. Clinical experts in PTSD were identified by consulting with the VA's National Center for PTSD and with VA Central Office mental health leaders. Members of the expert panel determined what key questions to ask providers related to the documentation of PTSD severity and its clinical course. Currently, cognitive interviews and focus group discussions with clinicians in the field are being conducted at five CHIR sites.

Parallel to this deductive process we are using statistical text mining techniques as a supplemental approach that permits the discovery of potentially useful text fragments as candidate clinical terms. Statistical text mining 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 have eluded the clinical team. At the very least it will provide confirmation of the importance of concepts identified though more traditional methods. In this paper, we report on our preliminary results and discuss plans to further refine the analysis.

METHODS

Data Preparation

To create an appropriate database for statistical text mining, a data table of patient notes labeled as being from a patient with PTSD or not needed to be constructed.

Through the administrative data available within the VA, a cohort of 5,165 unique OEF/OIF veterans who received care at the James A. Haley Veterans Hospital in Tampa Florida during FY 2007 (October 2006 to September 2007) were identified.

From this group additional administrative data was utilized to determine if the veteran had PTSD.

- Veteran must have been identified as having service connected disability with a diagnosis of PTSD by the VA Compensation and Benefits program.
- Veteran must also have had at least two outpatient visits where the diagnosis was PTSD (ICD-9-CM code 309.81)
- Veteran must be identified within the VistA EMR system problem list as having PTSD.

With this criterion in mind, 405 veterans were identified as being positive for PTSD.

A control group of 392 veterans was also created from the initial cohort. The control group was made up of veterans who were not identified in any of the PTSD criterion but were identified as having mental health disorders containing depression or anxiety. Using this definition, the diagnosis of PTSD was removed yet the control group still contained some type of mental disorder. This controls for terms that were universally prevalent in mental health jargon and at the same time highlight the terms that were utilized in the diagnosis of PTSD only.

Patient notes from the cohort of 797 (405+392) were gathered from the VA administrative data. This resulted in over 25,000 notes covering over 300 note types. Table 1 represents a subset of those note types.

ADDENDUM	PASTORAL CARE	PAIN MEDICINE CONSULT	OCCUPATIONAL THERAPY CONSULT
PSYCHIATRY OUTPAT		ADMISSION EVALUATION	PAIN MEDICINE OUTPAT CONSULT
PSYCHOLOGY OUTPAT	PSYCHIATRY E & M INTERDISCIPLINARY	AUDIOLOGY CONSULT	SPEECH PATHOLOGY CONSULT
MENTAL HEALTH OUTPAT E & M	EDUCATION DISCHRG	MENTAL HEALTH NURSING E & M	UROLOGY CONSULT
NURSING OUTPAT	NUTRITION INPATIENT	MENTAL HEALTH TREATMENT PLAN	CWT CONSULT
MENTAL HEALTH OUTPAT	RECREATIONAL THERAPY	NEUROPSYCHOLOGY CONSULT	OB GYN OUTPAT
PRIMARY CARE	ORTHOTICS PROSTHETICS CONSULT	PODIATRY CONSULT	OPHTHALMOLOGY NURSING
NURSING	ADVANCE DIRECTIVE DISCUSSION	VOCATIONAL REHABILITATION CONSULT	OPHTHALMOLOGY OUTPAT
PSYCHOLOGY GROUP CNSLG	MHICM	DENTISTRY CONSULT	PULMONARY OUTPAT
PRIMARY CARE NURSING	NURSING EMERGENCY DEPT TRIAGE	GASTROENTEROLOGY	SOCIAL WORK DOMICILIARY
MENTAL HEALTH OUTPAT GROUP CNSLG	NURSING OUTPAT MEDICATION MGT	PHARMACY OUTPAT	SOCIAL WORK NURSING LONG TERM CARE
PSYCHOLOGY	AGENT ORANGE PROGRAM	PHYSICAL MEDICINE REHAB E & M	SOCIAL WORK TELEPHONE ENCNTR
PSYCHIATRY	CONSENT	PSYCHIATRY INPATIENT E & M	UROLOGY OUTPAT
TELEPHONE ENCNTR	NEUROPSYCHOLOGY	DERMATOLOGY	VOCATIONAL REHABILITATION DOMICILIARY
PRIMARY CARE E & M	PRIMARY CARE MEDICATION MGT	NURSING TREATMENT PLAN	ADVANCE DIRECTIVE
PHARMACY OUTPAT MEDICATION MGT	PRIMARY CARE TELEPHONE ENCNTR	PHARMACY CONSULT	CASE MANAGER
SATP	NO SHOW	PHYSICAL THERAPY CONSULT	GEC
MENTAL HEALTH OUTPAT CONSULT	SPEECH PATHOLOGY	SATP CONSULT	LPN
NURSING FLOWSHEET	WOMENS HEALTH	WOMENS HEALTH NURSING	MENTAL HEALTH DISCHRG
MENTAL HEALTH GROUP CNSLG	NURSING TELEPHONE ENCNTR	NEUROLOGY CONSULT	NUTRITION GROUP CNSLG
PSYCHIATRY INPATIENT	DENTISTRY	NURSING OUTPAT TRANSFER SUMMARIZATION	PHARMACY MEDICATION MGT
PRIMARY CARE NURSING TRIAGE	VOCATIONAL REHABILITATION	PSYCHIATRY OUTPAT E & M	PHYSICAL THERAPY INITIAL EVALUATION
PSYCHOLOGY CONSULT	MENTAL HEALTH CONSULT	DENTISTRY OUTPAT	PRE OPERATIVE E & M
PRIMARY CARE H & P	PRIMARY CARE INITIAL EVALUATION	MENTAL HEALTH TELEMEDICINE	PULMONARY CONSULT
PRIMARY CARE ADMINISTRATIVE	HOMELESS PROGRAM	OPHTHALMOLOGY CONSULT	SMOKING CESSATION OUTPAT
H & P	PSYCHIATRY CONSULT	ORTHOPEDIC SURGERY OUTPAT	SPINAL CORD INJURY ANNUAL EVALUATION
MENTAL HEALTH TELEPHONE ENCNTR	EMERGENCY DEPT	PHARMACY OUTPAT CONSULT	GASTROENTEROLOGY NURSING
CWT	NEUROLOGY	RESPIRATORY THERAPY EDUCATION	MENTAL HEALTH ADMINISTRATIVE
EDUCATION	NURSING IMMUNIZATION	AUDIOLOGY	OB GYN
OCCUPATIONAL THERAPY	ORTHOPEDIC SURGERY	DENTISTRY ADMINISTRATIVE	OPTOMETRY CONSULT
SOCIAL WORK OUTPAT	IMMUNIZATION	GASTROENTEROLOGY OUTPAT CONSULT	OUTREACH
ADMINISTRATIVE	NURSING EMERGENCY DEPT DISCHRG	NURSING INPATIENT E & M	PASTORAL CARE INITIAL EVALUATION
C & P EXAMINATION	PSYCHIATRY ATTENDING	NUTRITION EDUCATION	SOCIAL WORK GROUP CNSLG
MENTAL HEALTH EDUCATION	MENTAL HEALTH CASE MANAGER	OEF/OIF CASE MANAGER	SPINAL CORD INJURY ATTENDING
NURSING ADMISSION EVALUATION	PODIATRY	PODIATRY OUTPAT	SPINAL CORD INJURY NURSING
SOCIAL WORK	POLYTRAUMA	ANESTHESIOLOGY PRE OPERATIVE E & M	SURGERY CONSULT
PHYSICAL THERAPY OUTPAT	PHYSICAL THERAPY OUTPAT CONSULT	DENTISTRY PROCEDURE	ASI
MENTAL HEALTH CNSLG	DOMICILIARY GROUP CNSLG	DERMATOLOGY CONSULT	CARE COORD HOME TELEHEALTH SUMMARY
PHYSICAL MEDICINE REHAB CONSULT	DOMICILIARY	NEUROLOGY OUTPAT	DERMATOLOGY OUTPAT
PHYSICAL THERAPY	CHIROPRACTIC	NONVA DIAGNOSTIC STUDY REPORT	INTERNAL MEDICINE ATTENDING INPATIENT
NURSING EMERGENCY DEPT E & M	PRIMARY CARE TRIAGE	PAIN MEDICINE OUTPAT E & M	INTERNAL MEDICINE
PRIMARY CARE NURSING E & M	SATP EDUCATION	PHYSICAL THERAPY OUTPAT DISCHRG	KINESIOTHERAPY INITIAL EVALUATION
SOCIAL WORK DISCHRG	SOCIAL WORK CONSULT	SOCIAL WORK INPATIENT	KINESIOTHERAPY
PHYSICAL MEDICINE REHAB	OPHTHALMOLOGY	NURSING CONSULT	NEUROLOGY OUTPAT CONSULT
PHYSICAL MEDICINE REHAB OUTPAT E & M	ORTHOPEDIC SURGERY CONSULT	NUTRITION CONSULT	NEUROLOGY PROCEDURE

Table 1. Examples of notes types – pre-filter

For data reduction purposes, the data was restricted to note types that were of mental health type origin. This was accomplished by instituting the following filters

For the **cases**,
- Note Title contains one of following terms either as a standalone word or embedded within a word
    - SUIC
    - SOCIAL
    - PSYCH
    - PTSD
    - MENTAL

OR the Note Title was ADDENDUM and the note contained one of the above terms

For the **controls**,
--Note Title contains one of following terms either as a standalone word or embedded within a word
    - SUIC
    - SOCIAL
    - PSYCH
    - MENTAL
    - DEPRES
    - ANX

OR the Note Title was ADDENDUM and the note contained one of the above terms

This filter resulted in a note reduction of approx 2/3^{rds}. After filtering, the new note count summed to 10,501; 9,197 case notes and 1,304 control notes. Additionally table 2 shows that this restricted the note type count to 48 from over 300.

PSYCHIATRY OUTPATIENT	MENTAL HEALTH CASE MANAGER
PSYCHOLOGY OUTPATIENT	SOCIAL WORK CONSULT
MENTAL HEALTH OUTPATIENT E & M	MENTAL HEALTH NURSING E & M
MENTAL HEALTH OUTPATIENT	NEUROPSYCHOLOGY CONSULT
ADDENDUM	MENTAL HEALTH TREATMENT PLAN
PSYCHOLOGY GROUP COUNSELING	PSYCHIATRY INPATIENT E & M
MENTAL HEALTH OUTPAT GROUP CNSLG	PSYCHIATRY OUTPATIENT E & M
PSYCHOLOGY	MENTAL HEALTH TELEMEDICINE
PSYCHIATRY	SOCIAL WORK INPATIENT
MENTAL HEALTH OUTPATIENT CONSULT	SOCIAL WORK TELEPHONE ENCOUNTER
MENTAL HEALTH GROUP COUNSELING	SOCIAL WORK NURSING LONG TERM CARE
PSYCHIATRY INPATIENT	SOCIAL WORK DOMICILIARY
PSYCHOLOGY CONSULT	MENTAL HEALTH DISCHARGE
MENTAL HEALTH TELEPHONE ENCOUNTER	SOCIAL WORK GROUP COUNSELING
SOCIAL WORK OUTPATIENT	MENTAL HEALTH ADMINISTRATIVE
MENTAL HEALTH EDUCATION	SOCIAL WORK OUTPATIENT CONSULT
SOCIAL WORK	PSYCHOLOGY TREATMENT PLAN
MENTAL HEALTH COUNSELING	MENTAL HEALTH E & M
SOCIAL WORK DISCHARGE	SOCIAL WORK INITIAL EVALUATION
PSYCHIATRY E & M INTERDISCIPLINARY	SOCIAL WORK CASE MANAGER
NEUROPSYCHOLOGY	MENTAL HEALTH TEAM CONSULT
MENTAL HEALTH CONSULT	PSYCHOLOGY DOMICILIARY
PSYCHIATRY CONSULT	SUICIDE RISK ASSESSMENT
PSYCHIATRY ATTENDING	SOCIAL WORK COUNSELING

All of the initial activity concerning scrubbing and filtering of data explained above was completed using SAS Enterprise Guide®. A combination of the prebuilt nodes and free code node was utilized.

Text Mining

Once the data was in the correct form it was imported into Text Miner and the notes were run through a series of models. Each model represented one of the possible weighting schemes available in Text Miner 4.1. There are three frequency weights (none, binary and log) and eight term weighting schemes (none, normal IDF, GF-IDF, entropy, chi-square, mutual information and information gain). This gave us a potential pool of 24 models. However the 'normal' term weight did not produce any viable results thus there were 21 models.

The data was initially broken into a 70%training/30%validation partition prior to the run of the models. This reduced the chance of over-fitting. Additional settings including utilizing the available stop lists and synonym lists within SAS, limiting the number of terms in each model to 500, dropping non-rolled up terms and ignoring different parts of speech. SVDs(singular value decomposition) were not computed.

Part of the desire was to visualize how this process could work in its simplest form. Therefore since there was no additional data cleansing other than what was explained above and no in-depth start/stop/synonym lists were utilized other than what was offered in the Text Miner package the total term count prior to analysis was large at around 50,000 terms. This original term list potentially includes misspelled terms and synonyms that possibly could have been corrected, rolled up or removed but were not.

Each model was analyzed two separate ways. First a statistical text mining analysis was performed followed by a regression analysis. The statistical analysis weighted each term individually based on the weight score in the above schemes. Each term was ranked and given a score based on the ranking. The regression analysis utilized stepwise regression. Again each term identified in the regression was ranked and scored based on the absolute value of the individual regression coefficient.

Scores were based on the maximum number of terms identified in any one model by either analysis. This turned out to be 113. Therefore the highest weighted or largest absolute value regression coefficient term within any particular model was given a score of 113 followed by the second highest weighted term getting a score of 112, the third highest 111 and etc. This scoring was done for both the weight schemes and the regression analysis. If a term appeared in one analysis but not the other the term would receive a score of 0 for the analysis that it did not appear in.

RESULTS

The resultant regression models proved to be highly predictive across the 21 models with sensitivity (0.975-0.983). However, specificity varied across the models (0.317-0.611). Models developed using the information gain term weight proved to have the highest specificity.

Table 3 shows the top 25 scoring terms resulting from the above analysis utilizing the term weights. The weighted scores produced a total of 343 distinct terms with 50 terms appearing in all 21 models. Terms that may have to do with mental health issues such as disorder, suicidal, and depression are evident

Table 4 shows the top 25 scoring terms resulting from the above analysis utilizing regression. The regression analysis provided 449 terms. In this analysis terms that have to do with mental health are more apparent such as anxiety, traumatic and depression which are all in the top three scoring terms.

Top 25 terms identified with Weight Term Method			Top 25 terms identified with Regression Method		
Term	**# of Models**	**Score**	**Term**	**# of Models**	**Score**
+ day	17	1893	+ anxiety	21	2906
+ year	17	1882	traumatic	18	2235
+ medication	16	1805	+ depression	18	2010
+ insight	16	1784	sc	15	1915
+ change	14	1770	+ goal	17	1912
+ veteran	17	1720	+ authorize	14	1875
+ disorder	14	1718	+ disorder	16	1755
+ pass	14	1680	+ well	15	1745
+ pain	15	1674	+ nightmare	14	1704
suicidal	16	1672	+ continue	18	1654
+ process	15	1643	+ back	18	1602
+ deny	14	1625	+ place	15	1600
+ discuss	14	1612	+ progress	12	1535
+ sleep	15	1597	irritability	12	1501
+ last	14	1575	+ father	16	1492
+ date	13	1565	+ evaluation	18	1448
+ affect	13	1558	+ impression	11	1432
+ month	13	1525	+ major	12	1423
+ continue	14	1522	ptsd	12	1418
+ experience	13	1517	+ psychotherapy	12	1407
+ review	14	1500	+ stress	12	1399
+ depression	12	1498	+ hallucination	11	1373
+ mood	14	1484	orientation	10	1361
+ contact	12	1445	+ quality	10	1335
+ drink	14	1444	oef	10	1307

Table 3. Examples of Highly Predictive Terms Identified Through Statistical Text Mining using Weighted Terms method

Table 4. Examples of Highly Predictive Terms Identified Through Statistical Text Mining using Regression method

Table 5 is a summation of each of the scores in each individual analysis.

Again we see some mental health terms at the top of the list such as ‘anxiety’, ‘depression’ and ‘disorder’. The average coefficient across all the models that the term was utilized in during the regression analysis is also available. Hence we see that ‘anxiety’, ‘depression’ and ‘disorder’ are associated with the anxiety and depression group while ‘PTSD’ is associated with the PTSD group as expected.

At this point it is unclear to us how to derive associations utilizing the weighted term method therefore if a term does not appear in any regression analysis models, the association is considered “unknown”(insight, sleep).

Top 50 terms identified – combined methods									
	Wt.	Reg.	Tot.	Avg. Coef		Wt.	Reg.	Tot.	Avg. Coef
+ anxiety	959	2906	3865	-3.72	sc	0	1915	1915	3.09
+ depression	1498	2010	3508	-1.64	+ mood	1484	425	1909	-0.08
+ disorder	1718	1755	3473	-2.86	+ experience	1517	377	1894	0.68
+ continue	1522	1654	3176	0.46	+ authorize	0	1875	1875	2.30
+ goal	1179	1912	3091	1.27	+ process	1643	216	1859	0.05
+ deny	1625	1239	2864	-0.45	+ affect	1558	299	1857	-4.39
+ medication	1805	649	2454	-4.19	+ place	221	1600	1821	1.64
+ month	1525	921	2446	-0.18	+ evaluation	362	1448	1810	-0.83
+ veteran	1720	694	2414	-0.13	suicidal	1672	121	1793	0.11
+ change	1770	547	2317	-0.19	+ pulse	1098	692	1790	0.73
+ session	1355	904	2259	1.22	+ insight	1784	0	1784	#N/A
+ no.	1074	1174	2248	-0.78	+ pain	1674	101	1775	0.15
traumatic	0	2235	2235	2.46	+ group	648	1121	1769	0.42
+ activity	1222	1012	2234	-0.64	+ stress	354	1399	1753	0.87
+ pass	1680	480	2160	0.86	+ well	0	1745	1745	-0.65
+ year	1882	268	2150	-3.94	+ review	1500	243	1743	-0.07
+ back	540	1602	2142	1.39	+ follow	1414	290	1704	0.13
ptsd	701	1418	2119	1.59	+ symptom	1394	297	1691	-4.96
+ nightmare	367	1704	2071	3.09	+ hallucination	313	1373	1686	1.10
+ counsel	1315	665	1980	-0.54	+ screen	1345	291	1636	-0.58
+ day	1893	80	1973	0.13	+ sleep	1597	0	1597	#N/A
+ assessment	1033	921	1954	-2.58	+ speech	1407	187	1594	4.14
+ high	651	1279	1930	-1.47	+ combat	475	1118	1593	1.06
axis	787	1138	1925	-1.56	+ present	1424	161	1585	3.98
+ discuss	1612	310	1922	0.20	+ history	883	701	1584	-0.65

Table 5. Top 50 combined scoring terms identified through Statistical Text Mining.

CONCLUSIONS

There are multiple potential benefits from this type of analysis. Using SAS Text Miner allows for the capability to process large number of document without the need to

- build an ontology
- have intensive programming skills
- have large amounts of resources (human capital, training, time)

Theoretically from this point a term list could now be shown to a panel of experts to determine accuracy. The term list is much smaller creating an environment where the clinician could, easier and in short order, determine inclusion/non inclusion status of each of the identified terms.

However there are also multiple issues with this type of analysis.

In addition to the mental health terms identified, in this particular process, we also see many non-mental health related terms. A more robust start/stop/synonym list should help address this. Building these lists are time consuming and resource intensive and part of the goal of this study was to see what kind of return we could get if investing as little capital as possible.

From what we understand, Text Miner 4.1 does not contain the capacity to rotate SVDs. This is a problem that SAS has addressed in Text Miner 4.2. Additionally, although some misspellings are captured, many are not. A more robust spell checker is now in place in Text Miner 4.2

The actual number of terms varied in each model, 34 at the low end and 113 at the high end. Utilizing a scoring system, as was done in this study, guaranteed each of the top weighted or scored terms would carry the same weight. However this also meant that the low weighted or scored terms could carry more weight in one model due to the lower number of terms. An option would be to score based on ranked percentage or a logarithmic style scoring system.

Going forward our plan is to repeat this analysis utilizing SAS Text Miner 4.2. Future analysis should also not just focus on single terms but also SVDs (multi-term topics) and potentially a combination of the two.

We will also be using this analysis for the MRSA data. With the MRSA data the logic to define a MRSA positive note is not as clear-cut as with PTSD data which will bring an added dimension into the analysis.

REFERENCES

[1] Frueh BC, Grubaugh AL, Elhai JD, Buckley TC. US Department of Veterans Affairs disability policies for posttraumatic stress disorder: administrative trends and implications for treatment, rehabilitation, and research. *Am J Public Health*. 2007 Dec;97(12):2143-5.

[2] Veterans with Post-Traumatic Stress Disorder (PTSD) Fact Sheet 2006.

[3] Cimiano P. Ontology learning and Population from Test, Algorithms, Evaluation and Applications. New York: Springer Science+Business Media, LLC. 2006.

[4] Ramchand R, Schell TL, Karney BR, Osilla KC, Burns RM, Caldarone LB, “Disparate prevalence estimates of PTSD among service members who served in Iraq and Afghanistan: Possible explanations,” Journal of Traumatic Stress, February 2010.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Matthew Richardson
HSR&D/RR&D Center of Excellence, Maximizing Rehabilitation Outcomes
James A. Haley VAMC (118M)
8900 Grand Oak Circle
Tampa, FL 33637-1022
Phone: 813-558-3975
E-mail: matthew.richardson@va.gov

This material is based upon work supported by the Office of Research and Development (add, as applicable: Medical Research Service, or Health Services Research and Development Service, or Rehabilitation Research and Development Service), Department of Veterans Affairs.