Journal of Evaluation in Clinical Practice
International Journal of Public Health Policy and Health Services Research

Journal of Evaluation in Clinical Practice ISSN 1365-2753

SHORT COMMUNICATION

Automated classification of psychotherapy note text: implications for quality assessment in PTSD care

Brian Shiner MD MPH,[1] Leonard W. D'Avolio PhD,[2] Thien N. Nguyen PhD,[3] Maha H. Zayed PhD,[4] Bradley V. Watts MD MPH[5] and Louis Fiore MD MPH[6]

[1]Attending Psychiatrist, VA Medical Center, White River Junction, VT, USA, Fellowship Director, VA-New England Healthcare Engineering Partnership, White River Junction, VT, USA and Assistant Professor, Dartmouth Medical School, Hanover, NH, USA
[2]Informatics Researcher, VA Boston Healthcare System, Boston, MA, USA, Associate Center Director for Biomedical Informatics, Massachusetts Veterans Epidemiology Research and Information Center, Boston, MA, USA, Assistant Professor, Center for Surgery and Public Health, Boston, MA USA and Assistant Professor, Harvard Medical School, Boston, MA, USA
[3]Research Computer Programmer, VA Boston Healthcare System, Boston, MA, USA and Lead Developer, Massachusetts Veterans Epidemiology Research and Information Center, Boston, MA, USA
[4]Clinical Research Psychologist, VA Medical Center, White River Junction, VT, USA
[5]Attending Psychiatrist, VA Medical Center, White River Junction, VT, USA, Associate Director, VA-New England Healthcare Engineering Partnership, White River Junction, VT, USA and Assistant Professor, Dartmouth Medical School, Hanover, NH, USA
[6]Attending Physician, VA Boston Healthcare System, Boston, MA, USA, Cooperative Studies Program Coordinating Center Director, Massachusetts Veterans Epidemiology Research and Information Center, Boston, MA, USA and Associate Professor, Boston University School of Public Health, Boston, MA, USA

Keywords
health care quality, access, and evaluation, natural language processing, psychotherapy, quality assurance, health care, quality of health care, stress disorders, post-traumatic

Correspondence
Dr Brian Shiner
VA Medical Center
215 North Main Street 11Q
White River Junction, VT 05009
USA
E-mail: brian.shiner@va.gov

Human Subjects Review: This study was approved by the Dartmouth College Committee for the Protection of Human Subjects (CPHS #21631).

Accepted for publication: 23 December 2010

doi:10.1111/j.1365-2753.2011.01634.x

Introduction

In recent years, studies have attempted to use various methods to characterize the quality of care for post-traumatic stress disorder (PTSD) delivered in United States Veterans Administration (VA) outpatient clinics. Dieperink *et al*. used manual chart review to characterize care for 150 veterans at three VA medical centres during the 2001 fiscal year [1]. They found wide variation in the types of social services, psychotherapy and pharmacotherapy received in the 6 months following entry into specialized PTSD programmes. For example, clinics in Minneapolis and Memphis tended to provide pharmacotherapy while clinics in Boston tended to provide psychotherapy. As a result, there was wide variation in the amount of care received; veterans in Minneapolis received an average of seven mental health contacts per year (visits with a psychiatrist, psychologist, social worker or other mental health practitioner) while veterans in Boston received an average of 16 mental health contacts per year. Although it was not clear which approach was superior, it was clear that PTSD care was not standardized among the three VA facilities. In order to include more sites and be able to generalize across sites, further studies on the quality of care for PTSD in the VA have attempted to use national administrative data rather than chart review to capture information about the process of care. This work has relied on a combination of Current Procedural Technology (CPT) codes, International Classification of Diseases, Ninth Revision (ICD-9) codes, and pharmacy data.

Cully *et al*. used administrative data to examine the receipt of psychotherapy in the VA nationally for the 12 months following initial diagnoses of PTSD, anxiety and depression in the 2004 fiscal year [2]. While the 77 743 veterans with new diagnoses of PTSD had a higher chance of receiving psychotherapy than veterans with anxiety or depression, the amount of psychotherapy was still very low; only 10% received an adequate number of sessions (defined as eight in this study), and the median wait to start psychotherapy was 50 days. Two studies using similar methods were published in 2010. Using stricter inclusion criteria, Spoont *et al*. examined care for 20 284 veterans with a new diagnosis of PTSD from the 2004 mid-fiscal year through the 2005 mid-fiscal year [3]. They evaluated whether veterans received a large enough medication supply that they could have gotten an adequate trial of pharmacotherapy or whether they received enough psychotherapy visits that they could have received an adequate trial of psychotherapy (defined, again, as eight visits). Based on this resource utilization, they concluded that, at most 33%, of veterans could have received an adequate trial of evidence-based treatment for PTSD. Seal *et al*. examined all mental health visits over the year following a new PTSD diagnosis in veterans returning from Iraq and Afghanistan [4]. The study included 49 425 veterans enrolling in VA care from the 2002 mid-fiscal year through the 2008 mid-fiscal year. They asserted that

delivery of evidence-based psychotherapies endorsed in the VA Mental Health Uniformed Services Package [5] (prolonged exposure [6] and cognitive processing therapy [7]) required at least nine sessions over 15 weeks and found that only 9.5% received this level of service. A key limitation in these three studies is that they tell us only the best possible scenario about the amount of psychotherapy that could have been delivered based on the number of visits – we do not know whether veterans actually received psychotherapy during these visits. Therefore, it is possible that these studies overestimate the amount of psychotherapy actually delivered to veterans with PTSD.

Researchers and policy makers wishing to understand care delivery for PTSD are left with a dilemma. Manual medical record review can generate detailed information about clinical processes, including psychotherapists' reports of the specific techniques they used in a session. However, the method is time-consuming and difficult to apply on a large scale. Administrative review techniques are applicable on a large scale, but are limited in the granularity of the information they provide. We learn how much of a given service practitioners report providing, but because we do not read the notes, we have little information about the content of those services.

Automated text-based information retrieval technologies, such as natural language processing (NLP) have the potential to bridge this gap by extracting detailed information found in a medical record review on the larger scale permitted by administrative review. NLP is an effort to have computers draw specific information from free text. The application of NLP has traditionally been limited by the need to customize programming for each new application. However, modern NLP applications can use a technique known as machine learning to 'teach' a computer to recognize patterns in documents [8,9]. Through the recognition of language patterns within a document the NLP application can help users make inferences about the content of the text.

We sought to understand whether using administrative data to determine the number of psychotherapy sessions veterans receive is equivalent to manual medical records review. We thought it was possible that psychotherapy billing codes might sometimes be misapplied to other services delivered by psychotherapy-oriented practitioners, such as psychologists and social workers. These might include intakes, psychological testing and case management. Alternatively, administrative data review might be accurate, making manual or automated review of note text an unnecessary method. Our primary hypothesis was that administrative data overestimates the number of psychotherapy sessions delivered to veterans when compared to manual chart review (as some sessions administratively coded as psychotherapy are actually used for other purposes). Our secondary hypothesis was that if administrative data review was inaccurate, our manual medical record review could be approximated using an automated NLP programme, creating the potential for a more accurate method to be efficiently applied to large-scale treatment studies.

Methods

Manual coding

We examined 6 months of routine clinical care for veterans in a single VA clinic who met symptomatic criteria for PTSD using the PTSD Checklist [10] from 2005 to 2007. We classified mental health notes using the manual chart review protocol designed by Dieperink *et al*. [1] during the 6 months following the initial assessment. Two psychologists reviewed and classified each note (e.g. medication management, individual psychotherapy, group psychotherapy, case management, other mental health service) and a psychiatrist acted as adjudicator in resolving differences during group coding sessions. For 100 consecutive subjects, we identified all notes associated with encounters that were administratively coded as individual psychotherapy without medication management as defined by Cully *et al*. ([2]; CPT codes 90804, 90806, 90808, 90810, 90812, 90814, 90845, 90875, 90876 and 96152). Through this process we identified 221 notes. We pasted each note into a text file and marked the file with a '1' or a '0' to indicate whether the manual coding team had determined that the note met Dieperink *et al*.'s criteria to be called individual psychotherapy.

Automated coding

We loaded the 221 notes into the Automated Retrieval Console (ARC), a VA-developed software application initially piloted to extract information from cancer-related pathology and imaging reports [11]. ARC was designed to eliminate the need to develop custom software coding or rules, a process that requires expert programmers that has generally limited the widespread implementation of text-based information retrieval technologies. Instead, ARC 'learns' from a set of gold standard interpretations made with manual medical record review (called annotation). ARC capitalizes on existing open source software to derive robust feature sets from NLP pipelines [12] and to classify these features using supervised learning [13]. The two classification models used in this study are Conditional Random Fields (CRFs) and Maximum Entropy (MaxEnt). To determine an appropriate model, ARC automatically iterates through various combinations of features and classifiers, evaluating the performance of each using 10-fold cross validation and the supplied reference set (in this case, 221 documents). ARC is described in greater detail elsewhere [11] and can be downloaded along with HTML and video tutorials at http://research.maveric.org/mig/arc.html.

Statistical methods

For our primary hypothesis, to compare administrative coding to manual coding (the gold standard in this study), we calculated the percentage of notes associated with a CPT code indicative of individual psychotherapy that were coded as individual psychotherapy by the manual rating team. For our secondary hypothesis, to compare manual to automated chart review, we measured performance in terms of recall (akin to sensitivity; fraction of the documents that are relevant to the query that are successfully retrieved), precision (akin to specificity; fraction of retrieved documents that are relevant to the search) and harmonic mean (F-measure; akin to receiver operating characteristic). These measures are defined in Table 1.

Results

The coding team was able to agree on a classification of all notes through the dual coding and adjudication process. Of the 221 notes

Table 1 Statistical measures used in information retrieval

Measure	Definition
Precision	|{retrieved documents} ∩ {relevant documents}| ÷ |{retrieved documents}|
Recall	|{relevant documents} ∩ {retrieved documents}| ÷ |{relevant documents}|
F-measure	2 × |(precision × recall) ÷ (precision + recall)|

{} = 'set of', ∩ = 'intersection': common elements between two documents sets.

Table 2 Top scoring combinations

Feature types	Recall	Precision	F-measure
Maximum entropy			
Canonical token	0.98	0.89	0.93
Canonical token, base token	0.98	0.89	0.93
Token	0.96	0.90	0.93
Conditional random fields			
Token	0.97	0.90	0.93
Token, base token	0.97	0.90	0.93
Token, named entity type	0.97	0.90	0.93

associated with an encounter administratively coded as individual psychotherapy, 126 (57%) were manually coded as individual psychotherapy. The remaining notes were generally intakes, psychological testing and case management notes.

We were able to replicate the manual raters' coding very well using ARC. The top scoring features using two different classification algorithms (MaxEnt and CRFs) are shown in Table 2. Tokens appeared consistently as valuable features for classification. MaxEnt proved to be slightly better suited for the task than CRFs. With over 52 combinations of classifiers and feature types attempted, top performance of several configurations converged at a recall of 0.97, a precision of 0.90 and a harmonic mean of 0.93.

Discussion

We confirmed our primary hypothesis: it appears that using counts of administrative codes overestimates the amount of psychotherapy delivered to veterans with PTSD. In our small sample, almost half of the encounters that would have been counted as the provision of psychotherapy in large administrative studies appeared to be records of services other than psychotherapy. While these services are valuable and were delivered in good faith to the benefit of American veterans, it may be inaccurate to count them as the provision of psychotherapy. Of course, our finding is based upon observations at a single site and is therefore not generalizable as a corrective estimate for multi-site studies of treatment provision. However, manual medical record review is labour-intensive and would not be feasible in the study of a national treatment system, such as the VA. Fortunately, in confirming our secondary hypothesis, we have established a potential method for raising the accuracy of automated data review to that of manual review, permitting the efficient performance of more accurate large-scale national treatment studies. Despite the linguistic complexities of understanding psychotherapy notes, ARC performed as well in this application as it did in the seemingly more straightforward application of classifying cancer-related pathology reports and better than it did in classifying cancer-related imaging reports [11].

More work is needed before we can use ARC as a mental health services research tool. The application of ARC in a larger multisite dataset would be necessary to demonstrate generalizability of the method. Furthermore, with the demonstrated efficacy of specific psychotherapeutic modalities for PTSD [6,7], it has become important to the VA that veterans not only receive psychotherapy, but that they receive specific types of psychotherapy [5]. ARC may hold potential in the automated classification of multiple subtypes of psychotherapy, including cognitive processing therapy and prolonged exposure. Such work would be important in order to understand the effect of the VA's evidence-based practice dissemination efforts.

This study suggests a potential limitation in current studies of the quality of care for PTSD in the VA, namely inaccurate coding of psychotherapy notes. In addition, we have also identified a potential tool to ameliorate this problem, use of NLP to automate the coding of psychotherapy notes based on the note text. These methods have the potential to take our understanding of care delivery in the VA further, by helping us understand not just whether psychotherapy occurred, but what type of therapy was done.

Acknowledgements

Dr Shiner's time is supported by the VA-New England Early Career Development Award Program. This work is made possible by funding from the Veterans Administration Cooperative Studies Program (CSP) and Health Services Research & Development through the Consortium for Healthcare Informatics Research (CHIR).

References

1. Dieperink, M., Erbes, C., Leskela, J., Kaloupek, D., Farrer, M. K., Fisher, L. & Wolf, E. (2005) Comparison of treatment for posttraumatic stress disorder among three Department of Veterans Affairs medical centers. *Military Medicine*, 170 (4), 305–308.
2. Cully, J. A., Tolpin, L., Henderson, L., Jimenez, D., Kunik, M. E. & Peterson, L. A. (2008) Psychotherapy in the veterans health administration: missed opportunities? *Psychological Services*, 5 (4), 320–331.
3. Spoont, M. R., Murdoch, M., Hodges, J. & Nugent, S. (2010) Treatment receipt by veterans after a PTSD diagnosis in PTSD, mental health, or general medical clinics. *Psychiatric Services*, 61 (1), 58–63.
4. Seal, K. H., Maguen, S., Cohen, B., Gima, K. S., Metzler, T. J., Bertenthal, D. & Mamar, C. R. (2010) VA mental health services utilization in Iraq and Afghanistan veterans in the first year of receiving new mental health diagnoses. *Journal of Traumatic Stress*, 23 (1), 5–16.
5. Kussman, M. J. (2008) VHA Handbook 1160.01: Uniform Mental Helath Services in VA Medical Centers and Clinics. Available at: http://www1.va.gov/vhapublications/ViewPublication.asp?pub_ID=1762 (last accessed 1 March 2010).
6. Foa, E. B., Hembree, E. A., Cahill, S. P., Rauch, S. A., Riggs, D. S., Feeny, N. C. & Yadin, E. (2005) Randomized trial of prolonged exposure for posttraumatic stress disorder with and without cognitive restructuring: outcome at academic and community clinics. *Journal of Consulting and Clinical Psychology*, 73 (5), 953–964.

7. Monson, C. M., Schnurr, P. P., Resick, P. A., Friedman, M. J., Young-Xu, Y. & Stevens, S. P. (2006) Cognitive processing therapy for veterans with military-related posttraumatic stress disorder. *Journal of Consulting and Clinical Psychology*, 74 (5), 898–907.
8. Uzuner, O., Goldstein, I., Luo, Y. & Kohane, I. (2008) Identifying patient smoking status from medical discharge records. *Journal of the American Medical Informatics Association*, 15 (1), 14–24.
9. Uzuner, O. (2009) Recognizing obesity and comorbidities in sparse data. *Journal of the American Medical Informatics Association*, 16 (4), 561–570.
10. Weathers, F. W., Litz, B. T., Herman, D. S., Huska, J. A. & Keane, T. M. (1993) The PTSD Checklist (PCL): Reliability, validity, and diagnostic utility. Paper presented at the *9th annual meeting of the ISTSS*, San Antonio, TX.
11. D'Avolio, L. W., Nguyen, T. M., Farwell, W. R., Chen, Y., Fitzmeyer, F., Harris, O. M. & Fiore, L. D. (2010) Evaluation of a generalizable approach to clinical information retrieval using the automated retrieval console (ARC). *Journal of the American Medical Informatics Association*, 17 (4), 375–382.
12. Ferrucci, D. (2004) UIMA: an architectural approach to unstructured information processing in the corporate research environment. *Natural Language Engineering*, 10, 327–348.
13. McCallum, A. K. (2002) MALLET: a machine learning for language toolkit. Available at: http://mallet.cs.umass.edu/ (last accessed 11 November 2010).