HHS Public Access

Author manuscript

Med Biol Eng Comput. Author manuscript; available in PMC 2019 July 01.

Published in final edited form as:

Med Biol Eng Comput. 2018 July ; 56(7): 1285–1292. doi:10.1007/s11517-017-1772-1.

Classifying Clinical Notes with Pain Assessment using Machine Learning

Samah Jamal Fodeh, PhD,
Department of Emergency Medicine, Yale Center of Medical Informatics, Yale University School of Medicine, New Haven, CT 06519-1315

Dezon Finch, PhD,
HSR&D Center of Innovation on Disability and Rehabilitation Research, James A. Haley Veterans Hospital, and the University of South Florida College of Public Health, Department of Health Policy and Management, 8900 Grand Oak Circle, Tampa, FL, USA 33637

Lina Bouayad, PhD,
HSR&D Center of Innovation on Disability and Rehabilitation Research, James A. Haley Veterans Hospital, and the University of South Florida College of Public Health, Department of Health Policy and Management, 8900 Grand Oak Circle, Tampa, FL, USA 33637

Stephen L. Luther, PhD,
HSR&D Center of Innovation on Disability and Rehabilitation Research, James A. Haley Veterans Hospital, and the University of South Florida College of Public Health, Department of Health Policy and Management, 8900 Grand Oak Circle, Tampa, FL, USA 33637

Han Ling, PhD,
Internal Medicine (Geriatrics), Yale University, 300 George St, New Haven, CT 06511

Robert D. Kerns, PhD, and
Departments of Psychiatry, Neurology and Psychology, Yale School of Medicine, and Pain Research, Informatics, Medical comorbidities and Education (PRIME) Center, VA Connecticut Healthcare System, West Haven, CT 06516.

Cynthia Brandt, MD MPH
Pain Research, Informatics, Medical comorbidities and Education (PRIME) Center, VA Connecticut Healthcare System, and Yale Center of Medical Informatics, Department of Emergency Medicine, Yale University School of Medicine, New Haven, CT 06519-1315

Abstract

Objective—Pain is a significant public health problem, affecting millions of people in the United States. Evidence has highlighted that patients with chronic pain often suffer from deficits in pain care quality (PCQ) including pain assessment, treatment and re-assessment. Currently, there is no intelligent and reliable approach to identify PCQ indicators inelectronic health records (EHR).

Corresponding author: Samah Jamal Fodeh, PhD. samah.fodeh@yale.edu, Department of Emergency Medicine, Suite 264F, Yale University, New Haven, CT 06519-1315, Phone:+1 (203) 298-4924, Fax:+1 (203) 785-4580.

Conflict of interest: The authors declare that they have no conflict of interest.

Hereby, we used unstructured text narratives in the EHR to derive pain assessment in clinical notes for patients with chronic pain.

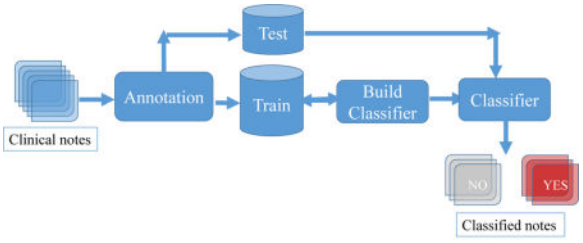
Materials and Methods—Our dataset includes patients with documented pain intensity rating ratings >= 4 and initial musculoskeletal diagnoses (MSD) captured by (ICD-9-CM codes) in fiscal year 2011 and a minimal one year of follow-up (follow-up period is 3-yr maximum); with complete data on key demographic variables. A total of 92 patients with 1058 notes was used. First, we manually annotated qualifiers and descriptors of pain assessment using the annotation schema that we previously developed. Second, we developed a reliable classifier for indicators of pain assessment in clinical note.

Results—Based on our annotation schema, we found variations in documenting the subclasses of pain assessment. In positive notes, providers mostly documented assessment of pain site (67%) and intensity of pain (57%), followed by persistence (32%). In only 27% of positive notes did providers document a presumed etiology for the pain complaint or diagnosis. Documentation of patients' reports of factors that aggravate pain was only present in 11% of positive notes. Random Forest classifier achieved the best performance labeling clinical notes with pain assessment information, compared to other classifiers.; 94%, 95%, 94%, 94% was observed in terms of accuracy, PPV, F1-score, and AUC, respectively.

Discussion—Despite the wide spectrum of research that utilizes machine learning in many clinical applications, none explored using these methods for pain assessment research. In addition, previous studies using large datasets to detect and analyze characteristics of patients with various types of pain have relied exclusively on billing and coded data as the main source of information. This study, in contrast, harnessed unstructured narrative text data from the EHR to detect pain assessment clinical notes.

Conclusion—We developed a Random Forest classifier to identify clinical notes with pain assessment information. Compared to other classifiers, ours achieved the best results in most of the reported metrics.

Graphical abstract



Framework for detecting pain assessment in clinical notes

Keywords

pain; assessment; classification; Random Forest

INTRODUCTION

Pain is a significant public health problem, affecting an estimated 100 million Americans at an annual cost of up to US$635 billion in medical treatment and lost productivity.[1] Evidence has highlighted that patients with chronic pain often suffer from deficits in Pain Care Quality (PCQ).[1] PCQ indicators assess the degree to which providers and/or healthcare systems follow evidence-based practice standards and accepted standards of care. Once the presence of pain is recognized, the following PCQ processes should be followed and documented in patients' electronic health records: a timely and appropriate comprehensive pain assessment, development and enactment of an integrated pain treatment plan informed by the assessment, and ongoing monitoring and reassessment of the effectiveness of the plan.[2,3] Efforts to improve PCQ hinge on the identification of reliable PCQ indicators and promotion of their use in systematic quality improvement efforts.

Currently, there is no intelligent and reliable approach to identify patients with clinically significant pain in the electronic health records (EHR).[4] Capturing data elements related to pain in most EHRs is not standardized. For example, diagnostic codes of pain are not uniquely identified;[4,5] and the patient self-reported pain rating scale[6] is only modestly accurate in identifying patients with pain.[7] A study by Geotzke et al[8] attempted to identify potential patients with chronic pain. This work, however, was criticized by Tain et al[4] due to poor evaluation, complexity and unfeasibility in primary care setting.[4] These authors proposed an algorithm, with good performance, to identify patients with chronic pain using opioid prescriptions in addition to pain diagnostic codes and pain intensity ratings in the EHR. Maeng and colleagues used variables such as number of encounters in the study period, insurance information, follow-up time, and opioid prescriptions to detect high cost pain patients.[9] In another study on patients with chest pain that remained undiagnosed six months after first presentation, Jordan and colleagues used structured data including age at index presentation, sex, body mass index (BMI), smoking status, neighborhood deprivation, prescriptions for lipid lowering drugs, and specific comorbidities to detect pain.[10] Other than pharmacological and procedure based interventions in which specific, easily retrievable codes are used to document care, it is difficult to capture pain assessment or key aspects of integrated care plans and their enactment. Moreover, while a growing number of pain-related studies have utilized structured and easily retrievable coded fields in the EHR, limited research has explored the utility of provider narratives in clinical notes. Bui and Zeng[11] extracted snippets of text from clinical notes that contain the word "pain", then built a classifier to categorize the notes with "pain" or "no pain". This approach is greatly limited, however, and encourages future efforts to build a more comprehensive classifier system.

In this study, we leveraged unstructured text narratives in the EHR to derive PCQ indicators. We focused on detecting indicators of a comprehensive pain assessment in clinical notes for patients with chronic pain. Other than routine documentation of pain intensity ratings and results of diagnostic tests and procedures, pain assessment is buried in narratives written by health care providers and nurses.[2] Assessment of pain has many facets including intensity, quality, persistence, diurnal variation, aggravating and alleviating factors and reports of pain interference with physical and emotional functioning.[2,12] In prior work, we demonstrated the capacity to reliably annotate and extract documentation of these indicators from primary

care provider narrative progress notes.[2] In subsequent work, we demonstrated responsivity to change for several of these variables in the context of a system-level effort to improve PCQ in the primary care setting.[13] The manual annotation, however, is extremely time intensive and effortful, and for some indicators, reliability remained relatively low despite considerable effort. Thus, the use of automated systems offers important opportunities to a more reliable and efficient approach to capturing pain assessment in the EHR.

We employed machine learning (ML) algorithms to analyze unstructured narrative text data in the EHR to develop a reliable classifier that detects pain assessment in clinical notes. Identifying notes with pain assessment is an important step towards developing decision-support tools to enable health care providers to deliver the best possible care to patients with chronic pain. The proposed framework of our study is illustrated in Figure 1.

MATERIALS and METHODS

All materials were derived from the Department of Veterans Affairs (VA) EHR. The clinical notes were obtained from the Veterans Health Information Systems and Technology Architecture (VistA) EHR through the VA Informatics and Computing Infrastructure (VINCI). We included patients with documented pain intensity rating score >= 4,[14] and initial musculoskeletal diagnoses (MSD) captured by (ICD-9-CM codes) in fiscal year 2011 and a minimal one year of follow-up (follow-up period is 3-yr maximum); with complete data on key demographic variables. A total of 9,940 patients were selected from 130 VA facilities including 8,268 males and 1,672 females. A total of 376,487 clinical notes with 2,172 distinct types/note titles were associated with the patients' encounters. The set of clinical notes was further reduced using primary care clinic (stop) codes 322,323,350 which cover the "full range" of services delivered in clinics. 323- Primary Care Clinic which captures the core of clinical services. 322- Women's Clinic this stop code captures the women veterans' primary care clinical services. 350: Geriatric Primary Care: Each geriatric encounter must be recorded with stop code 350.[15] This reduction retained 138,274 notes with 849 types. Narrowing the notes by type to include only primary care physicians' notes left 99,481 clinical notes with 101 types. These clinical notes are written by primary care providers and follow the subjective, objective, assessment plan (SOAP) format.[16,17] For the purposes of our analysis, we sampled 1058 clinical notes that belong to 92 males and females with mean ages 68 and 58, respectively (keeping all notes of a unique patient). We then divided these notes into 10 sets to control the workflow of the annotation process. That is, once the annotators completed annotating one set, the agreement statistics and adjudication are examined on that set. Lessons learned from annotating this set are then reflected on future sets.

Annotation

We developed annotation schema, based on our previous work,[2,13] similar to guidelines developed to support traditional chart review but included more explicit details about the specific text strings that should be coded. The schema is a computer program that allows the annotator to highlight specific text strings and assign a classification label to the string. We used eHost to create the schema.[18] To measure inter-annotator agreement (IAA), a first

training set of 50 clinical notes was annotated by each of the annotators. Agreement statistics were calculated on the training notes as a measure of reliability. Differences between annotators were adjudicated by a third expert on the team to help ensure that the annotations were valid and reliable. This is the standard practice for these kinds of text studies based on the well documented and well cited references.[19–23] Adjustments to guidelines were made and a second set of training notes was annotated and the IAA statistics was 65%. The agreement statistics IAA were calculated in eHOST which uses a simple agreement statistic and F-Measure (harmonic mean of precision and recall). If any given annotation by an annotator is matched by an annotation with the same class, it was deemed a match. Matches were calculated for each annotator compared to the other. Final agreement is reported as a percent agreement between the two annotators broken down for each class. Annotated spans of text are recorded with their start and stop offsets identifying their exact position in the text. Two annotations with overlapping spans and assigned the same class are considered a match. Any differences in the spans are resolved during the adjudication process. The goal was to annotate each instance of a mention of pain assessment. Pain assessment could appear in clinical notes in different forms or sub-classes. These subclasses were derived from published policy guidance and recognized standards in the field regarding the key components of a comprehensive pain assessment.[2] In addition, pain management experts were consulted to help operationalize these subclasses, and further operationalization occurred during the development phase to improve reliability of the coding. The included pain assessment subclasses are:

1. Pain Mention: The presence or absence of pain as experienced by the patient. Note that words such as “aching”, “discomfort”, “tenderness” and “discomfort”, among similar words, are all considered pain mentions.
2. Intensity: Patient reported description of the intensity of pain including numeric pain intensity ratings or verbal descriptions of the pain severity (i.e., mild, moderate, severe).
3. Quality: Description of the nature or character of the pain such as sharp, dull, nagging, burning, electric, and so forth.
4. Persistence: The degree to which the pain continues or persists such as “constant, relentless, always present” versus “intermittent, comes and goes, fluctuates.”
5. Diurnal variation: Description of the high and low levels of pain intensity as they occur at specific times of the day. Examples include “worse in the morning” and “achy all day.”
6. Aggravating factors: Factors associated with worse pain such as “standing more than 5 minutes” and “when I’m alone.”
7. Alleviating factors: Factors associated with less pain, most often behaviors and activities such as “hot showers” and distraction such as “reading” or “talking on the telephone”.
8. Functional assessment: The manner and degree to which pain interferes with patients’ lives including interference with physical or emotional functioning such

as “because of pain cannot work, wakes up at night” and “uses wheelchair because of pain.”

9. Pain Etiology: Medical conditions that include the word pain such as “myofascial pain syndrome” or “chronic low back pain” among many others as well as conditions presumed to be painful including a wide range of pain-related conditions such as degenerative joint disease (DJD), osteoarthritis (OA), and injury, among many others. The word pain in the phrase is NOT to be double annotated as pain mention.

10. Pain Site: The location of the pain (e.g. knee, lower back, joint). Note that, as a rule, no pain site is to be annotated without a pain mention.

11. Pain Related Diagnostics: The results of any test or consult to diagnose the pain condition such as radiographic findings or results of electromyographic evaluation, or results of a consult with physical therapy, neurology or other specialists who assess and treat pain conditions. In this sub-class, findings need to be related to a potential source of pain but not necessarily to pain mention.

The presence of any sub-class in a note indicated that the corresponding patient had been assessed for that dimension of pain. While annotating, the phrase, in a note, indicative of a particular sub-class was selected. In other words, a binary approach to coding the presence or not of each specific sub-class or dimension was employed. Pain assessment may be mentioned multiple times within a note indicative of one or multiple sub-classes. We accounted for assertions and negations in the text. An average of 72% (median 71%) IAA was obtained on the 10 sets of clinical notes.

Generating the reference standard

We generated a reference standard dataset for pain assessment that we used to build the classification system. To label clinical notes for pain assessment, we used the extracted annotations. Within a note, single or multiple sub-classes of pain assessment might be present. If a note had at least one annotation of a sub-class of pain assessment, then it was deemed positive and labeled as “Yes.” In the negative case where no pain assessment annotations were detected in a note, the label was “No.” Out of the 1058 notes in our sample, 596 were positive for the presence of documentation of at least one pain assessment sub-class (56%; labeled “Yes”) compared to 462 negative notes.

Pain assessment classification

In our experiments, we used the scikit-learn machine learning toolkit[24] to build the classifiers. Data representation was the first task towards building the classifier. We used the Natural Language Toolkit (NLTK)[25] to extract words/features from each clinical note in the sample. In particular, the natural language processing (NLP) pipeline included tokenization, stop words removal and stemming (using porter stemmer which is supported by the NLTK implementation). We used scikit-learn machine learning toolkit[24] to generate the bag-of-words (BOW) representation of the notes. The extracted words via NLTK were passed on to Scikit-learn to compute corresponding frequencies. Each note was represented via a feature vector wherein features (words in the notes) are weighted using their frequencies. The

classifier was built using the BOW representation of notes. We experimented with several classifiers to classify each note as either pain assessment or not. We observed performance and selected the classifier with best results. The list of classifiers included:

- *K-Nearest neighbor (KNN)*[26,27]: Classification is computed from a simple majority vote of the nearest neighbors of each point: a query point is assigned the data class which has the most representatives within the nearest neighbors of the point. The K is a parameter that represents the number of neighbors of the query point. We experimented with many values (4,6,8,10) and found that K=6 gave best results. In the scikit-learn implementation we used the Euclidean distance to compute the distance between a query point and its neighbors.
- *Decision tree (DT)*[28,29]: Non-parametric supervised learning method that creates a model that predicts the value of a target variable by learning simple decision rules inferred from the data features. To measure the quality of the split we used the Gini impurity[30] which measures misclassifications of the split. We allowed the tree to grow to its maximum depth i.e. until leaves are pure or they contain at least one clinical note.
- *Support Vector Machine (SVM)*: An SVM classifier learns the region of feature space (i.e. combinations of features) that corresponds to a “pain assessment” note. The boundary of this region is known as the decision boundary. SVM maps each feature vector to a decision value, which is the signed distance of the note to the decision boundary. For example, decision values of 5, 1, −1 would correspond to high confidence of assessment, a less certain of assessment, and no assessment in the note, respectively. We used the support vector classification (SVC) implementation in scikit-learn to do the classification. We trained the classifier with the radius basis function (RBF) Kernel for which two parameters must be determined C and Gamma. C trades off misclassified clinical notes and the complexity of the decision boundary. Low C produces a smooth decision boundary while high C aims to classify more clinical notes correctly. Gamma defines the importance and influence of a clinical note.[31,32] We set the parameters to their default values as recommended by scikit-learn so C=1 and Gamma=1/number of features.
- *Random Forest (RF)*:[33,34] An ensemble learner that combines the predictions of several base classifiers built with a given learning algorithm in order to improve generalizability and robustness over a single classifier. Each tree in the RF is built from a sample drawn with replacement from the training set. When splitting a node during the construction of the tree, the best split among a random subset of the features is picked. The prediction of the ensemble RF classifier is given as the averaged prediction of the individual classifiers. To build the random forest we included 100 classifiers.

Evaluation

We split clinical notes into three equal portions where two-thirds of the data was utilized for training and one-third for testing. We performed 10-fold cross-validation[35–37] to train the

classification system. In cross-validation, the training set is split into 10 smaller sets or folds. The classification model is trained on k-1 folds and tested or evaluated on the remaining fold using any of the evaluation measures below. This process is repeated 10 times such that the model will be tested on each one of the folds. The average performance on the 10 folds will be reported for the classification system. Dietterich[38] and Bouckaert[39] studied the 10-folds of cross-validation and proposed other ways of generating the 10 samples or increasing the samples to 100 instead of 10, however, the standard method described above remains the most widely used validation procedure.[40] To evaluate performance, we used the following measures.

Positive Predictive Value (PPV): Precision—The ability of the classifier not to label as positive a sample that is negative and is the ratio TP/(TP+FP) where TP is the number of true positives and FP the number of false positives.

Sensitivity: Recall—The ability of the classifier to find all positive samples and defined as the ratio: TP/(TP+FN) where FN is the number of false negatives.

F1-score—*T*he harmonic mean of precision and recall, can be interpreted as a weighted average of the precision and recall and is computed as 2 * (precision * recall)/(precision + recall).

The best value for all three measures is 1 and the worst value is 0.

RESULTS

Annotation

We used the annotations to label clinical notes with the sub-classes of pain assessment. A positive note with pain assessment could have single or multiple sub-classes as labels. Table 1 shows the total number of unique sub-classes that appeared in the 596 positive notes (having pain assessment). If a sub-class appeared multiple times in a note, it was counted as one. Based on our data sample and evident form Table 1, the majority of clinician notes assessed one to four aspects of pain for patients. In 381 (65%) of the notes for which any pain assessment sub-class was present, patients were assessed for at least 2 to 4 sub-classes of pain. A smaller, percentage (23%) of the notes included assessments of 4 to 6 sub-classes, and 9% of the notes had evidence of 6 to 9 sub-classes of pain assessment. None of the notes included annotations of 10 or11 pain assessment sub-classes. It is important to mention that 26 out of 596 positive notes had treatment annotations related to pain but we do not include them in the analysis in this table because we are focused on pain assessment sub-classes (this explains the total of 570 notes in the table).

Figure 2 shows the distribution of sub-classes of pain assessment across the “positive” notes for which any documentation consistent with an acknowledgement of the presence of pain was present. The pain mention sub-class appeared in 94% of these notes (560 out of 596 positive notes of pain assessment). In the remaining 36 notes, the sub-class pain diagnostics was detected which is, by definition, not necessarily dependent on the pain mention sub-class (i.e., use of the word “pain” was not required). Providers mostly documented

assessment of pain site (67%) and intensity of pain (57%), followed by persistence (32%). In only 27% of positive notes did providers document a presumed etiology for the pain complaint or diagnosis. Documentation of patients' reports of factors that aggravate pain were documented approximately 11% of the positive notes. The remaining sub-classes of pain assessment, including assessment of pain quality, pain diagnostics, interference with functioning, factors that alleviate pain and diurnal variation in pain intensity were all documented in less than 10% of these positive notes.

Classifying clinical notes with pain assessment

We then applied three single classifiers including SVM, K-nearest neighbor, and decision tree as well as the random forest ensemble classifier to detect pain assessment in clinical notes. Table 2 summarizes the average performance of the different classifiers of ten runs on the training and test sets in all measures. On the training data, best performance in most measures was observed for RF where the highest accuracy and AUC were achieved, .93 and .92, respectively. Cross validation error was computed for all classifiers. SVM, RF and NN had about ~.02 error compared to .013 error of DT. On the training data DT showed close performance to RF. The best performance of RF is explained by the combined predictions of several base classifiers which likely improved generalizability and robustness over a single classifier. In our experiments, we combined 100 single learners for prediction. Despite the similar overall performance of DT and RF on the training set, On the test data RF wins all classifiers including DT. K-nearest neighbor, however, outperformed all classifiers in terms of sensitivity on the training and test sets; .93 and .95 respectively. We used six neighbors to estimate the label of a given clinical note (i.e. k=6); adding beyond that did not seem to change the results. KNN, however, did noticeably worse than all classifiers in terms of the other measures for both training and test data sets as shown in the table.

DISCUSSION

We developed an automated system based on ML to detect pain assessment in clinical notes. We founded an annotation schema that we used to extract information about pain assessment from clinical notes. We then generated a reference standard of labeled notes to build a classifier that can find patient encounters with documentation of pain assessment. Notes including pain assessment annotations were deemed positive examples and the remaining were assigned to the negative class. We experimented with multiple classifiers, among which random forest classifier achieved the best results.

There have been numerous studies that used ML to classify EHR clinical notes relative to a certain health care condition or outcome of interest.[41–49] Use cases included cancer related problems such as microcalcification and colon cancer, mental health and psychiatric diseases, falls in elderly patients; amongst many others. Despite the wide spectrum of this research, the use of intelligent machine-based methods for pain assessment research has not been explored. In addition, previous studies using large datasets to detect and analyze characteristics of patients with various types of pain have relied exclusively on the readily available billing and coded data as the main source of information.[50–53] This study, in

contrast, harnessed unstructured narrative text data from the EHR to detect pain assessment clinical notes; a focus that has not been explored before.

In this analysis, we made the case that structured and coded data are not sufficient to aid the research of pain assessment and thus PCQ. As we have shown in the annotation section, not all primary care clinical notes pertaining to patients with MSD diagnoses include useful information for pain; about half (56%) of the reference standard clinical notes (596/1058) had pain annotations. These findings are perhaps consistent with clinical observations that, although the sample was comprised of patients with known MSD who have reported pain of at least moderate severity during at least one encounter, pain may not always be a salient concern for these patients and their providers at all clinical encounters.[54] (add Goulet citation) These findings encourage a broader approach that takes into account unstructured clinical data when designing health related quality measures or making health-related conclusions or decisions. Our classification system is potentially useful for retrieving clinical notes with the focus interest of pain assessment from primary care; a more discerned set of notes for subsequent analysis.

The pain assessment classifier is potentially useful for health care providers, health services researchers, or other entities interested in improving PCQ. We envision it as an agent that can automatically sift through the EHR to pull clinical notes with pain assessment for further pain quality research and performance improvement initiatives. For example, in the context of a facility level initiative designed to successfully implement a stepped care model of pain management, we employed a manualized approach to assessing these dimensions of PCQ to monitor improvements in pain care in the primary care setting.[55] This project is an early step in developing an automated approach to extracting key pain-relevant information from clinical notes that is not otherwise available in structured data, for example, information about patient functioning, often considered an important outcome for pain clinical trials.[56] The importance of this system is amplified given the “Big Data” nature of the EHR. Unstructured text narratives in clinical notes is particularly characterized as “Big Data” due to their large volume and high dimensionality, thus very expensive and effortful to extract embedded useful information manually and automatically. Hereby, we propose to use our classification system to detect clinical notes of interest i.e. pain assessment and then use those for pain quality focused research. Our study has a limitation related to the specificity of the pain assessment information detected. Although the annotations in the reference standard included the different and specific sub-classes of pain assessment that exist in the notes, our classifier only determines if a clinical note includes components of a comprehensive pain assessment note or not. For example, the current system doesn’t generate information about the quality or quantity of these components (e.g., the extent of functional impairment, or the specific quality of patients’ pain experiences). Thus, our system represents a first step to detect indicators of providers’ documentation of key dimensions of a comprehensive pain assessment in a “Big Data” source.

CONCLUSION

The pain assessment classification system that we developed represents an important first step in developing an automated system that can potentially aid in improving PCQ. It

provides practical information useful to inform future policies and initiatives to improve the care of patients with pain. The classification system classifies clinical notes (in the test set) with pain assessment with an AUC and PPV equivalent to .94 and .95 respectively. Performance of RF ensemble classifier was compared to other single classifiers including SVM, KNN, and DT. In future work, we intend to build a more granular classifier that captures more specific and actionable information related to the different types of pain assessment in the notes. We plan to employ more sophisticated machine learning algorithms to build the new classifier.

Acknowledgments

Funding: This study was funded by NIH National Center for Complementary and Alternative Medicine – grant number (1R01AT008448-01). It was also partially funded by the Veterans Affairs.

Biography

Fodeh is a data miner and Dezon is an informatics researcher. Lina has a degree in data mining and Luther is a biostatistician. Kerns and Brandt are clinical experts.

References

1. Simon LS. Relieving pain in America: A blueprint for transforming prevention, care, education, and research. Journal of Pain & Palliative Care Pharmacotherapy. 2012; 26(2):197–198.
2. Dorflinger LM, Gilliam WP, Lee AW, Kerns RD. Development and application of an electronic health record information extraction tool to assess quality of pain management in primary care. Translational behavioral medicine. 2014; 4(2):184–189. [PubMed: 24904702]
3. Hooten, W., Timming, R., Belgrade, M. Assessment and management of chronic pain. Bloomington, MN: Institute for Clinical Systems Improvement; 2013. https
4. Tian TY, Zlateva I, Anderson DR. Using electronic health records data to identify patients with chronic pain in a primary care setting. Journal of the American Medical Informatics Association. 2013; 20(e2):e275–e280. [PubMed: 23904323]
5. Sinnott PL, Siroka AM, Shane AC, Trafton JA, Wagner TH. Identifying neck and back pain in administrative data: defining the right cohort. Spine. 2012; 37(10):860–874. [PubMed: 22127268]
6. Plaisance L. Pain—Clinical Manual. Home Healthcare Now. 2000; 18(8):556.
7. Krebs EE, Carey TS, Weinberger M. Accuracy of the pain numeric rating scale as a screening test in primary care. Journal of General Internal Medicine. 2007; 22(10):1453–1458. [PubMed: 17668269]
8. Goetzke, G., Johns, T., Reid, M., Borg, J., Carlson, A. Chronic pain patient identification system. Google Patents; 2001.
9. Maeng DD, Stewart WF, Yan X, et al. Use of electronic health records for early detection of high-cost, low back pain patients. Pain Research and Management. 2015; 20(5):234–240. [PubMed: 26291127]
10. Jordan KP, Timmis A, Croft P, et al. Prognosis of undiagnosed chest pain: linked electronic health record cohort study. bmj. 2017; 357:j1194. [PubMed: 28373173]
11. Bui DDA, Zeng-Treitler Q. Learning regular expressions for clinical text classification. Journal of the American Medical Informatics Association. 2014; 21(5):850–857. [PubMed: 24578357]
12. Sellinger, JJ., Wallio, SC., Clark, EA., Kerns, RD., Ebert, M., Kerns, R. Comprehensive pain assessment: The integration of biopsychosocial principles. Cambridge University Press; New York: 2010.
13. Anderson D, Zlateva I, Lee A, Tian T, Khatri K, Ruser CB. Stepped care model for pain management and quality of pain care in long-term opioid therapy. Journal of rehabilitation research and development. 2016; 53(1):137. [PubMed: 27006068]

14. Haskell SG, Brandt CA, Krebs EE, Skanderson M, Kerns RD, Goulet JL. Pain among veterans of operations enduring freedom and Iraqi freedom: Do women and men differ? Pain Medicine. 2009; 10(7):1167–1173. [PubMed: 19818028]
15. Affairs DoV. CHAPTER 264: PACT PRIMARY CARE CLINIC (PPCC). 2015
16. Weed LL. Medical records, patient care, and medical education. Irish Journal of Medical Science (1926–1967). 1964; 39(6):271–282.
17. Cameron S, Turtle-Song I. Learning to write case notes using the SOAP format. Journal of Counseling & Development. 2002; 80(3):286–292.
18. South BR, Shen S, Leng J, Forbush TB, DuVall SL, Chapman WW. A prototype tool set to support machine-assisted annotation. Paper presented at: Proceedings of the 2012 Workshop on Biomedical Natural Language Processing. 2012
19. Hripcsak G, Rothschild AS. Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association. 2005; 12(3):296–298. [PubMed: 15684123]
20. Hripcsak G, Heitjan DF. Measuring agreement in medical informatics reliability studies. Journal of biomedical informatics. 2002; 35(2):99–110. [PubMed: 12474424]
21. Ogren PV, Savova G, Buntrock JD, Chute CG. Building and evaluating annotated corpora for medical NLP systems. Paper presented at: AMIA Annual Symposium Proceedings. 2006
22. Ogren PV, Savova GK, Chute CG. Constructing evaluation corpora for automated clinical named entity recognition. Paper presented at: Medinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems. 2007
23. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008; 35(128):44.
24. SKI Learn. 2016. http://scikit-learn.org/stable/tutorial/index.html
25. Bird S. NLTK: the natural language toolkit. Paper presented at: Proceedings of the COLING/ACL on Interactive presentation sessions. 2006
26. Cunningham P, Delany SJ. k-Nearest neighbour classifiers. Multiple Classifier Systems. 2007; 34:1–17.
27. Peterson LE. K-nearest neighbor. Scholarpedia. 2009; 4(2):1883.
28. Quinlan JR. Induction of decision trees. Machine learning. 1986; 1(1):81–106.
29. Safavian SR, Landgrebe D. A survey of decision tree classifier methodology. IEEE transactions on systems, man, and cybernetics. 1991; 21(3):660–674.
30. Xia F, Zhang W, Li F, Yang Y. Ranking with decision tree. Knowledge and information systems. 2008; 17(3):381–395.
31. Gunn SR. Support vector machines for classification and regression. ISIS technical report. 1998; 14:85–86.
32. Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST). 2011; 2(3):27.
33. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. Journal of chemical information and computer sciences. 2003; 43(6):1947–1958. [PubMed: 14632445]
34. Liaw A, Wiener M. Classification and regression by randomForest. R news. 2002; 2(3):18–22.
35. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. Paper presented at: Ijcai. 1995
36. Refaeilzadeh, P., Tang, L., Liu, H. Cross-validation Encyclopedia of database systems. Springer; 2009. p. 532-538.
37. Zhang P. Model selection via multifold cross validation. The Annals of Statistics. 1993:299–313.
38. Dietterich TG. Approximate statistical tests for comparing supervised classification learning algorithms. Neural computation. 1998; 10(7):1895–1923. [PubMed: 9744903]
39. Bouckaert RR. Choosing between two learning algorithms based on calibrated tests. Paper presented at: Proceedings of the 20th International Conference on Machine Learning (ICML-03). 2003

40. Ross, KA. Cache-conscious query processing Encyclopedia of Database Systems. Springer; 2009. p. 301-304.
41. Fodeh SJ, Trentalange M, Allore HG, Gill TM, Brandt CA, Murphy TE. Baseline Cluster Membership Demonstrates Positive Associations with First Occurrence of Multiple Gerontologic Outcomes Over 10 Years. Experimental aging research. 2015; 41(2):177–192. [PubMed: 25724015]
42. Begg RK, Palaniswami M, Owen B. Support vector machines for automated gait classification. IEEE Transactions on Biomedical Engineering. 2005; 52(5):828–838. [PubMed: 15887532]
43. Widjaja E, Zheng W, Huang Z. Classification of colonic tissues using near-infrared Raman spectroscopy and support vector machines. International journal of oncology. 2008; 32(3):653–662. [PubMed: 18292943]
44. Orrù G, Pettersson-Yeo W, Marquand AF, Sartori G, Mechelli A. Using support vector machine to identify imaging biomarkers of neurological and psychiatric disease: a critical review. Neuroscience & Biobehavioral Reviews. 2012; 36(4):1140–1152. [PubMed: 22305994]
45. El-Naqa I, Yang Y, Wernick MN, Galatsanos NP, Nishikawa RM. A support vector machine approach for detection of microcalcifications. IEEE transactions on medical imaging. 2002; 21(12):1552–1563. [PubMed: 12588039]
46. Lee Y, Lee C-K. Classification of multiple cancer types by multicategory support vector machines using gene expression data. Bioinformatics. 2003; 19(9):1132–1139. [PubMed: 12801874]
47. Pakhomov SV, Hanson PL, Bjornsen SS, Smith SA. Automatic classification of foot examination findings using clinical notes and machine learning. Journal of the American Medical Informatics Association. 2008; 15(2):198–202. [PubMed: 18096902]
48. McCart JA, Berndt DJ, Finch D, Jarman J, Luther S. Using Statistical Text Mining to Identify Falls in VHA Ambulatory Care Data. Paper presented at: AMIA. 2012
49. Fodeh S, Benin A, Miller P, Lee K, Koss M, Brandt C. Laplacian SVM Based Feature Selection Improves Medical Event Reports Classification. Paper presented at: 2015 IEEE International Conference on Data Mining Workshop (ICDMW). 2015
50. Cicero TJ, Wong G, Tian Y, Lynskey M, Todorov A, Isenberg K. Co-morbidity and utilization of medical services by pain patients receiving opioid medications: data from an insurance claims database. PAIN®. 2009; 144(1):20–27. [PubMed: 19362417]
51. Breen AC, Carr E, Langworthy JE, Osmond C, Worswick L. Back pain outcomes in primary care following a practice improvement intervention:-a prospective cohort study. BMC musculoskeletal disorders. 2011; 12(1):1. [PubMed: 21199576]
52. Berger A, Sadosky A, Dukes E, Edelsberg J, Oster G. Clinical characteristics and patterns of healthcare utilization in patients with painful neuropathic disorders in UK general practice: a retrospective cohort study. BMC neurology. 2012; 12(1):1. [PubMed: 22289169]
53. Sullivan MD, Edlund MJ, Fan M-Y, DeVries A, Braden JB, Martin BC. Risks for possible and probable opioid misuse among recipients of chronic opioid therapy in commercial and medicaid insurance plans: The TROUP Study. Pain. 2010; 150(2):332–339. [PubMed: 20554392]
54. Goulet JL, Kerns RD, Bair M, et al. The musculoskeletal diagnosis cohort: examining pain and pain care among veterans. Pain. 2016; 157(8):1696–1703. [PubMed: 27023420]
55. Moore BA, Anderson D, Dorflinger L, Zlateva I, Lee A, Gilliam W, Tian T, Khatri K, Ruser C, Kerns RD. The stepped care model of pain management and quality of pain care in long-term opioid therapy. J Rehab Res Develop. 2016; 53(1):137–146.
56. Dworkin RH, Turk DC, Farrar JT, Haythornthwaite JA, Jensen MP, Katz NP, Kerns RD, et al. Core outcome measures for chronic pain clinical trials: IMMPACT recommendations. Pain. 2016; 113:9–19.

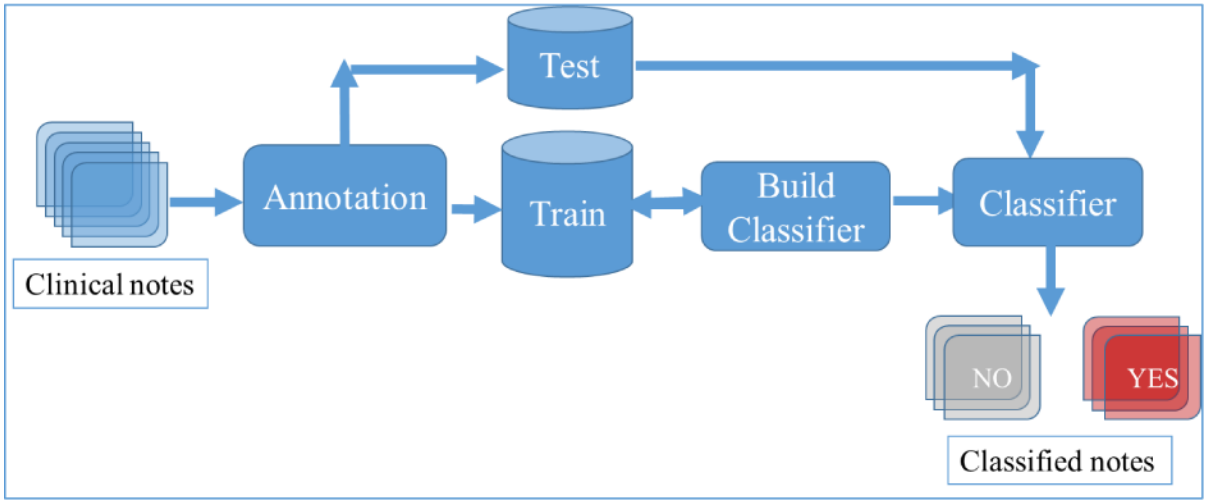


Figure 1.
Framework for detecting pain assessment in clinical notes.

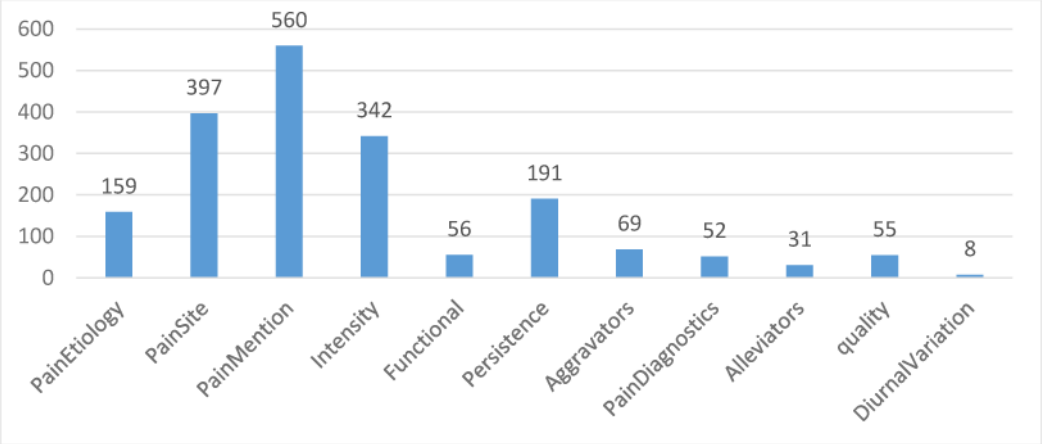


Figure 2.
frequency of different pain assessment sub-classes in positive notes (596)

Table 1

Number of different pain assessment sub-classes in clinical notes

Number of pain	Number	Percentage
Sub-classes ≤ 2	168	29%
2 < sub-classes≤4	213	36%
4 < sub-classes≤6	137	23%
6 < sub-classes≤9	52	9%

Table 2

Performance of classifiers

Training Data					
	Accuracy	**PPV**	**Sensitivity**	**F1-score**	**AUC**
SVM	0.88	0.86	0.92	0.87	0.88
RF	**0.93**	**0.91**	0.92	**0.92**	**0.92**
KNN	0.84	0.80	**0.93**	0.86	0.86
DT	0.91	0.90	0.91	0.91	0.91
Test Data					
SVM	0.89	0.87	0.94	0.90	0.90
RF	**0.94**	**0.95**	0.93	**0.94**	**0.94**
KNN	0.86	0.81	**0.95**	0.88	0.87
DT	0.92	0.93	0.92	0.93	0.92