Title

Phenotype Modeling and Outcome Mapping for Pain Management Decision Support

Principal Investigators and Team Members

PI: David Juckett, Ph.D., CTSI, Michigan State University
Team: Eric P. Kasten, Michigan State University
Philip L. Reed, Michigan State University
Fred N.Davis, ProCare Pain Solutions
Mark Gostine, Michigan Pain Consultants
Joseph Gardiner, Michigan State University
Rebecca Risko, ProCare Pain Solutions

Organization

Michigan State University, East Lansing, MI

Inclusive Dates of Project

August 1, 2014 to July 31, 2017

Federal Project Officer

Bryan B. Kim, PhD
Health Scientist / Program Officer
Division of Health Information Technology (IT)
Center for Evidence and Practice Improvement (CEPI)
Agency for Healthcare Research and Quality (AHRQ)

Acknowledgement of Agency Support

This project was supported by grant number R21HS022335 from the Agency for Healthcare Research and Quality. The content is solely the responsibility of the authors and does not necessarily represent the official views of the Agency for Healthcare Research and Quality.

Grant Award Number

R21HS022335

Abstract

Purpose:

The goal for this project was to show proof-of-principle that chronic-pain phenotypes could be determined from the medical data collected by a large community specialist practice (Michigan Pain Consultants (MPC)), and that these phenotypes could be linked to treatments and outcomes.

Scope:

To accomplish this purpose, the MPC electronic medical record (EMR) was evaluated to determine if the various components were suitable for phenotype construction. This required an examination of the elements within the MPC progress notes, their routine patient questionnaires, and their practice management data.

Methods:

Natural language processing techniques were used to extract concepts from the progress notes. An ontology suitable for community chronic pain medicine was constructed and an exemplar-based approach to extract the concepts for each patient was developed. The findings were rendered into feature vectors for latent class analysis. The 130 questions of the MPC patient questionnaire were consolidated into a 14 biopsychosocial feature vectors using factor analysis. Proof-of-concept phenotypes were generated from the feature vectors with preliminary latent class analysis on approximately 10,000 patient records.

Results:

The following results are recapped:

a) Data repository and annotation metrics.
b) Natural language processing algorithm approach based on exemplars.
c) Algorithm evaluation results.
d) Patient medication extractions.
e) Comorbidities identified in progress notes.
f) Clustering of patients into proof-of-concept phenotype groups.
g) Patient pain reporting veracity.
h) Longitudinal data overview.
i) Fuzzy Classification of Concepts Using Machine Learning.

Future publications and work are outlined.

Key Words:

Chronic pain, Phenotypes, Natural language processing, Progress notes, Patient-reported outcomes

Purpose

Specific Aims

The overall goals for our project are to isolate person-in-chronic-pain phenotypes from the data collected by a large community specialist practice (Michigan Pain Consultants (MPC)), link them to treatments and outcomes, and to use these results to create a model that can serve as the basis for future construction of a clinical decision support engine capable of enhancing patient outcomes. The much narrower aims for this AHRQ R21 grant were to evaluate the data from the MPC electronic medical record (EMR) and determine if it can be used in patient phenotyping and, if so, to show proof-of-principle results for various analysis components needed to convert the data to phenotypes. The two aims are:

1. Identify, extract, and organize data to support phenotype-intervention-outcome model construction. This requires analysis of the structured and unstructured data from the MPC clinics, which is composed of administrative data, dictated and formatted progress notes, and regular patient reported outcomes captured by a questionnaire with over 130 questions probing biopsychosocial patient attributes. A major component of this aim is to use natural language processing techniques to extract pertinent information from the progress note narratives so that it can be united with the structured data.

2. Iteratively construct and evaluate pain phenotype-intervention-outcome models until optimized to available data. This aim involves various combinations of factor analysis, cluster analysis, and structural equation modeling to build the model(s) that will be predictive of the phenotype – best treatment – best outcome axis.

Scope

The scope of this project is captured in the milestone table from this R21 grant application, which is reproduced in Table 1 together with brief comments on status. The comments indicate the progress achieved at each stage and the alterations made as the project developed.

When we created these milestones, we had great hope that this grant would be sufficient to create models capable of supporting clinical decision support tools. While we made great progress with Aim 1 (extracting data), the modeling (Aim 2) continues to remain a challenge because this grant was insufficient to extract all the phenotype features from the data (see Figure 1 for an overview), thus preventing us from generating the type of model we sought. We also now realize that we must expand our vision to create more comprehensive models that encompass longitudinal patient trajectories and longitudinal phenotypes. With that now in mind, the last milestone (iterative fine-tuning of models) is not be an appropriate endpoint. We now believe that this task must wait until the models can incorporate the time trajectories of patients undergoing the repetitive experiences of diagnosis, treatment, improvement, relapse, return to clinic, diagnosis, … and so on.

Table 1. Project milestones from grant application and status comments.

Milestones	Status Comments
Aim 1: Identify, extract, and organize data to support phenotype-intervention-outcome model construction	
Assemble NLP pipelines	We have designed and built an exemplar-based NLP pipeline
Train pipeline components	Our exemplar approach is 'trained' on human annotator-derived text instances linked to ontology classes
Deploy linkage analyzer	We have successfully implemented the linkages provided by the relationship exemplars to the output array for each document
Catalog extracted contents	We have generated ~450 Million algorithm-produced annotations and have stored them in a combination of a nested file structure and various Python dictionaries that link patients to annotation features
Aim 2: Iteratively construct and evaluate pain phenotype-intervention-outcome models until optimized to available data	
Outline initial phenotype model	We have learned enough to realize that our phenotype model must eventually be based on longitudinal patient trajectories because chronic pain is managed, not cured. For this R21 grant we created the beginnings of a static phenotype, proof-of-principle model based on latent factor and latent class analyses.
Utilize data to begin model-building	We have published a factor analysis[1] revealing potential phenotype factors from the PHA patient questionnaire.
Compare analysis methods	Global comparisons cannot be done yet. However extraction of each feature group was optimized by comparing methodologies for mapping to ontologies.
Incorporate results of Aim 1	Pain, treatments, relief, drugs, comorbidities, and negation have been incorporated to date
Iterative model refinement	May not be a good use of time for static phenotypes.

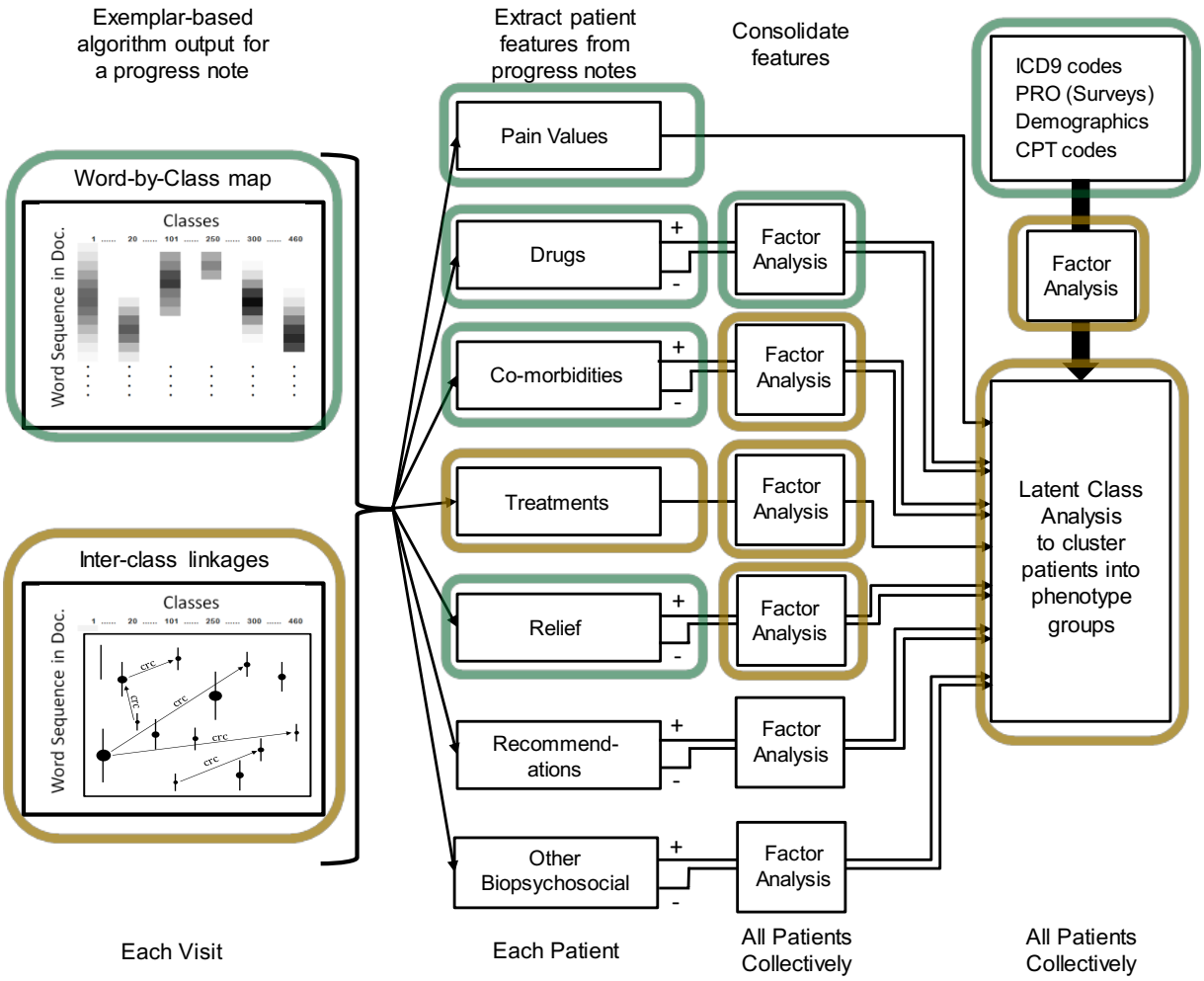

Figure 1. Schematic overview of project goals spanning from each visit to each patient to all patients. The grant has allowed a proof of concept for several steps in this project. Those have been highlighted: Green indicates finished and orange indicates partially finished

Methods and Results

1. Reports on various project components

a). Data metrics

In the submission proposal for this grant, a table of metrics was presented showing the number annotations generated by human annotators on a stratified random sample of progress notes. That information is expanded here in Table 2 by showing more details on the number of NLP annotations generated by our analysis and the number of patients with multiple visits for the period under study, (2010 – June 2014). To summarize, approximately 19,000 patients, of the 113,000 on record, visit the clinics an average of five times per year. These patients are

predominantly middle age and above. Associated with these visits are coding data, progress notes dictations, plus patient-reported status and outcomes using ProCare's Pain Health Assessment (PHA™) survey tool. The structured data is readily accessible to analysis, but to make complete use of the full data collection the contents of the progress note dictations require concept extraction and analysis using NLP technology, which is described in the next section. Initial NLP-generated annotations, as shown in the table, are approximately 450 million. These constitute a rich resource for further study.

Table 2. Metrics of available MPC patient data. Each line constitutes its own topic. Quantity abbreviations: K ≡ thousand; M ≡ million. Values are approximate.

Category	Value	Category	Value	Category	Value
Patients on record	113K	**Females**	67K	**Males**	46K
Mean Age (sd)	58.6 (17.5)	**Females**	59.4 (18)	**Males**	57.4 (16.6)
Visits (2010-14)	442K	**10+ Visits**†	30K	**100+ Visits**†	1K
Prog. Notes	288K	**10+ Notes**†	10K	**30+ Notes**†	800
Notes – Initial NLP	244K	**Sentences**	5.4M	**Algorithm Annotations**	450M
PHA questionnaires	84K	**iPHA**‡	23K	**cPHA**‡	61K
CPT codes	2.8M	**Avg per patient**	31		
ICD9-CM	1.5M	**Avg per patient**	17		

† - patients with this number of visits or progress notes

‡ **-** iPHA is intake PHA; cPHA is cumulative (follow-up) PHA

During the period 2015-2017, additional task-specific annotations were collected by student annotators, as mentioned in the 2nd annual report. Table 3 gives a summary view of all the human annotations collected with this project.

Table 3: Metrics of human-generated annotations.

Metric	Value	Notes
As of July, 2013		
Classes[1]	147,857	Number of mentions (instances) in the text assigned to a node in the class hierarchy
Linking Attributes[2]	136,292	Number of links assigned between class mentions
Classes Used	80.2 (4.3) %	Mean (sd) percentage of the 460 classes used during the annotation of the 500 documents by the four annotators
Coverage	77.8 (8.7) %	Mean (sd) percentage of the documents' text assigned to an annotation class
2015		
Cross-check Annotations	234,328	Evaluation of algorithm class assignments to sentences. 13,785 sentences evaluated on 500 fresh document.
2016-17		
Progress note content evaluation	~30,000	Twenty categories identified on random sets of documents by three student annotators
Longitudinal series	~32,000	Approximate number of evaluations of note content for approximately 2600 notes representing longitudinal series on 200 patients
Confusion matrix annotations	~13,300	TP, TN, FP, FN evaluations of various classes assigned to sentences (297) identified as containing relief by the algorithm

1. A class taxonomy structure was used. It contained 13 top classes with multiple branching sub-classes. A total of 460 final leaf classes comprised the hierarchy tree. (Simple examples would include: Drug name, Drug dose, Drug schedule, Pain intensity, Outcome pain relief.)
2. Classes were linked with attributes selected from a list of 39 possibilites. (Examples include: has dose, has drug schedule, has quality, has effect, has target, has value, has location laterality)

b). NLP algorithm

As mentioned in earlier reports, we have diverted from our plans for natural language processing that were outlined in the grant application. We had anticipated using off-the-shelf components based on UIMA and cTAKES to create pipelines for document processing. It became clear early on that this was not the best approach. These tools are still being built by large communities of researchers and have architectures for continuous processing of documents generated by various medical workflows. Virtually all of the pipeline components would have needing training to MPC data and that would have required more time and money than was available. In virtually all implementations of these pipelines, the focus is on extracting specific concepts for specific needs. We needed a method that could broadly identify the concepts embodied by chronic pain patients and their treatments. Hence we built an ontology to

represent these concepts, annotated a sample of documents, and used an exemplar-based concept extraction algorithm to process the progress notes.

The algorithm outline is shown in Figure 2 and depicted pictorially in Figure 3. To summarize this approach, we use annotation instances generated by human annotators from a random sample of notes to generate exemplars for matching to text regions of previously unseen documents (see Table 3). Two annotation collections were created. The first was to a low resolution class set. The second was to a high resolution ontology. In the second case, interclass relationships among classes were also captured. Analysis of unseen documents entailed extracting overlapping n-gram word segments and converting them to bag-of-words and bag-of-bi-character constructs for comparison to similar constructs stored for all the annotation exemplars. The fuzzy alignment approach consisted of choosing the maximum Jaccard index scores among the comparisons and retrieving the class(es) associated with the optimal annotation exemplar. The results were collected into discourse-by-class arrays for each document with scores occupying the cells of the array. These are basically class maps of the document discourse. They form the bases for subsequent analyses.

The algorithm is part of a manuscript for submission to the Data and Knowledge Engineering. As part of the basic aspects of the concept extraction method, we show in the manuscript the advantages of the bag-of-bi-character fuzzy alignment approach, both for its added benefits over bag-of-words and its increased speed over approximate string matching algorithms. We also show that the annotation collections were likely to be representative samples of the concepts in the documents, as determined by Zipf plot linearity and slope. We present the ROC curves for the method, which was requested by previous reviewers of a longer form of the manuscript submitted to the Journal of Biomedical Informatics.

The ROC curve was generated by using the confusion matrix values (TP, TN, FP, and FN) in a 5x5 cross-validation on the 500 annotated notes and thresholds applied to the values of the output array on these documents. As the thresholds increase in value, the false positive rates decreased, followed by decreases in the true positive rates, sketching out a locus of 2-D histogram intensities that define the ROC. The mean values were fit with the beta distribution for each of the 25 cases in the 5x5 cross-validation. In Figure 4a, the ROC is calculated from 72 parent classes in the chronic pain ontology, where classes related grammar and quantities were removed, and several child class trees were collapsed into one parent. The area under the ROC curve (AUC) when including all classes was 0.71 (not shown). With consolidation, the AUC improved to 0.90.

Most of the progress notes from MPC's EMR are constructed similarly and exhibit an expanded form of the general SOAP note (Subjective Objective Assessment Plan) although some variations among physicians occur routinely. We examined the ability of the exemplar algorithm to assign the top level class groups to the appropriate sections of the progress notes. The individual word usage through the text of the notes is not sufficiently unique to reliably distinguish the conceptually different regions (not shown). Rather, it requires phrases to best align regions with classes. In Figure 4b, the regions of the document dominated by the various top level classes of the taxonomy are shown (vertical scale). The solid circles are mean locations identified by the annotators, while the open triangles are those predicted by the algorithm. The error bars are the standard deviations around the peak locations for the occurrences of classes within the document. There are no significant differences between the human annotator identified positons in the document and those obtained by the algorithm. This provided another

test that the algorithm is extracting concepts from the progress notes suitable for further analysis.

Figure 2. Outline of analysis steps

Analysis Algorithms

1. *Basic Algorithm: Fuzzy matching*
- 1.1 Read in document
- 1.2 Separate into sentences with NLTK and convert to all lower case
- 1.3 For each sentence:
  - 1.3.1 Remove punctuation, determiners, and "dr", "ms", "mr", "mrs" leave all other stop words and symbols
  - 1.3.2 Extract overlapping word *n-gram*s of length 1,2,3,4,5 for each sentence Convert each *n-gram* to bag of words (BoW)
  - 1.3.3 Remove spaces from *n-gram*s and convert to bag of overlapping bi-characters (BoB)
- 1.4 For each *n-gram* BoW and BoB
  - 1.4.1 Calculate Jaccard Index (Eqn 1) similarity to each dictionary exemplar BoW and BoB using only those stored exemplars with sizes n to 4n in word length
  - 1.4.2 Select maximum Jaccard Index between the two bag types for an exemplar
  - 1.4.3 Sort Jaccard index values across all dictionary exemplars
  - 1.4.4 Capture the *n* top Jaccard index values and store with *n-gram* and scores; where, score is given by Equation 1, *n* is the same as *n-gram* length, and results are linked to word positions in document
- 1.5 Store all classes and scores linked to each word in the sentence

2. *Post-processing A:* Create discourse (word stream) by class output array for each document
- 2.1 Rows are identified by sentence number, word number in document, and word string
- 2.2 Columns are identified by class
- 2.3 Cells for each [word, class] intersection consist of the sum of all scores for that word derived for that class from all the *n-gram*s containing that word, (Eqn 3)

3. *Post-processing B*: Fit normal distributions to regions of class vectors in output array
- 3.1 From each document's output array:
  - 3.1.1 For each sentence:
    - 3.1.1.1 For each class vector
      - 3.1.1.2 Smooth twice with 3-point binomial kernel
      - 3.1.1.3 Locate peaks after smoothing
      - 3.1.1.4 Use number of peaks as number of normal distributions needed
      - 3.1.1.5 Fit smoothed class vector to sum of normal distributions
      - 3.1.1.6 Store means, standard deviations, and amplitudes for all identified peaks Store in a nested dictionary by document, sentence, and class.

4. *Post-processing C*: Use predicted inter-class relationships fine-tune output array scores
- 4.1 For each sentence. For each class.
- 4.2 Get mean locations of peak scores from normal fits (above)
- 4.3 Examine all class(a)-slot-class(b) possibilities documented in human annotation set
  - 4.3.1 If class(b) also exists in sentence, determine distance to class(a)
  - 4.3.2 Compare to distances documented in human annotation set
  - 4.3.3 Weight new score by differences in distances and frequency of class-slot-class (Eqn 4)

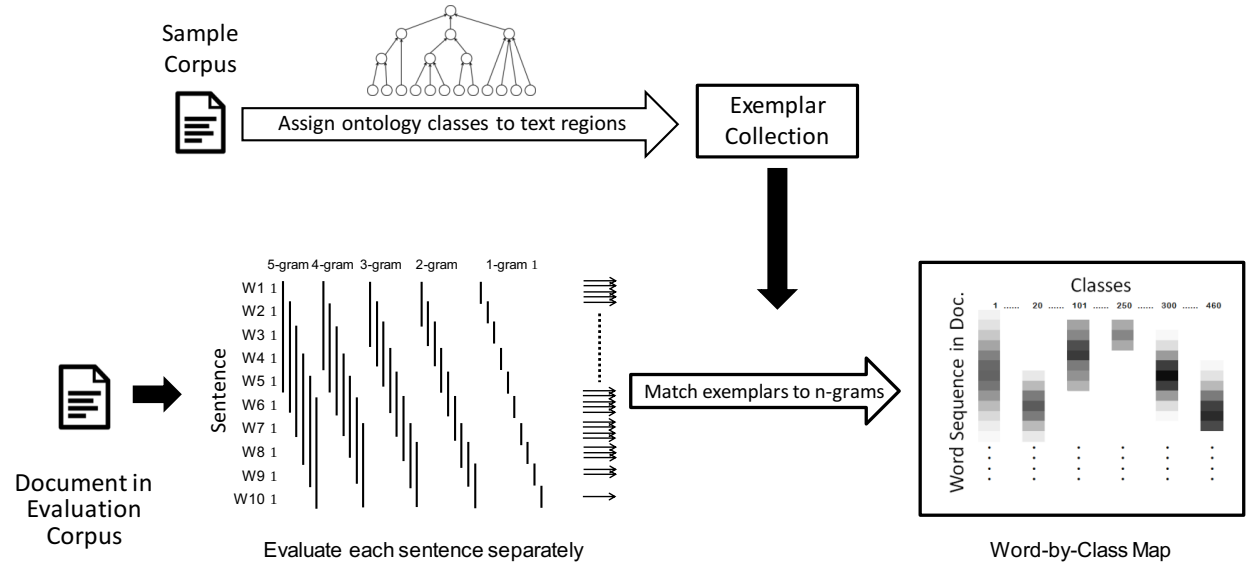

Figure 3. Flow diagram of the steps described in the basic algorithm and the first post-processing stage described in Methods and in Figure 1. This highlights that concept extraction from progress notes is linked to a chronic pain ontology.

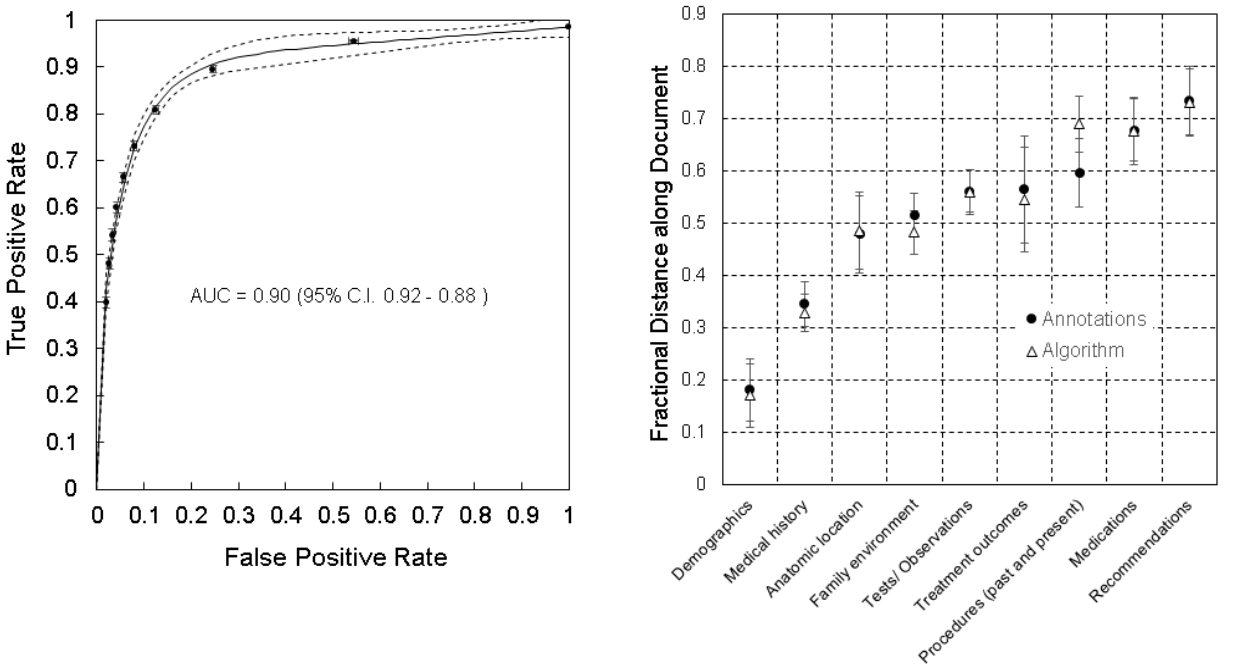

Figure 4. a) Results of 5x5 cross-validation for 72 class groupings. Mean (solid circles) and 95% confidence intervals for beta distribution fits to the 2D histograms. **b)** The positional distribution of score peaks for several top parent classes of the ontology. Human annotated documents (solid circles) and the algorithm output (open triangles) are shown for the same 500 document sample. Each document is normalized to a word length of 1.0

c). Algorithm results

In the 2nd annual report, we described two use cases to demonstrate the utility of the exemplar-based, fuzzy alignment method using either the low resolution classes or the high resolution ontology. In both cases, regions of the progress notes were aligned with concepts embodied in the class structure. Using the low resolution classes, we showed that we can capture the canonical structure of the progress notes. Using the high resolution ontology, we extracted pain relief, treatment modality, and polarity.

From the high resolution case some conclusions were readily apparent. Medications were generally not as effective as steroid injections or surgery. Surgery, however, had a higher percentage of negative comments in the patient reported outcomes than steroid injections. The highest percentage of negative comments occurred for medications, which appeared to indicate pain medications were not as effective as desired. This supports preliminary observations extracted from the PHA by Dr. Mark Gostine of MPC regarding the effectiveness of opioids[2].

We also showed that the relief estimates extracted from the progress notes show the same distribution as the answers to a single question of the PHA which asks the percentage of relief experienced by the patient. This distribution has the same skewed shape as the distributions for steroid injections and surgery.

In another exploration of the results, we sampled those documents associated with clinic visits where patients also filled out the pain health assessment (PHA) questionnaire (38,278 visits) (See Juckett et al.[1] for background on this PRO instrument.). The algorithm predicted that 90.6% of the documents contained at least one sentence addressing some degree of pain relief. Examining a random draw of 100 notes from this group, two human annotators identified 91 documents with relief statements. The algorithm predicted only a slightly different subset of documents. The recall, precision, and F1 scores between human annotators and the algorithm was 0.90, 0.93, and 0.91 respectively. The *kappa* score for inter-annotator agreement was 0.79, although we consider the kappa value a weak indicator of agreement because of the imbalance between true positives and false positives, as well as the large number of true negatives in the analysis leads to unrealistic values for agreement by chance, as discussed by Feinstein & Cicchetti, 1990[3].

For the same sample of 100 documents, 397 sentences were identified as possibly containing relief concepts. Within these 397 sentences, the intervention classes of injections, prescription drugs, and surgeries were also labeled by the algorithm along with polarity and scope. Human annotators labeled the relief, interventions, and polarity assignments with either true positive (*TP*) or false positive (*FP*) to enable calculation of precision. We did not attempt to evaluate recall because this was a sample chosen to have relief statements and, therefore, already biased toward high recall. Since relief is often associated with treatments, the classes related to injections, drugs, and surgeries were also biased to high recall. The precision values are shown Table 4, and were calculated using $\Sigma \Psi_{TP}/(\Sigma \Psi_{TP}+\Sigma \Psi_{FP})$, where Ψ represents the scores at the word-x-class intersection of the output array (see Figures 1 & 3). The Ψ_{TP} and Ψ_{TP} values were the maximal Ψ values assigned to the phrases in question, while the assignment to either TP or FP was determined by the annotators.

Table 4. Precision estimates for algorithmic identification of various concepts in a sample of 397 sentences in 100 documents, as validated by human annotators.

	Precision Mean (sd)
Relief	0.83 *(0.05)*[a]
Injections	0.93 *(0.05)*
Drugs	0.93 *(0.05)*
Surgeries	0.72 (0.15)
Polarity	0.97[b]

a). Standard deviation (sd) was a measure of the variation between annotators.
b). Polarity was sufficiently robust that only one annotator was used to evaluate

d). Medication used by patients and assignments to categories.

Named entity recognition of medications in the progress notes and prescriptions issued by MPC revealed 663,585 drug mentions within 77,900 progress notes and the associated visit prescription records of the EMR. These included over-the-counter drugs, those prescribed by other providers, as well those prescribed by MPC. They fell into approximately 769 specific drug types that could be consolidated into 23 general categories given by: Analgesic non-opioid; analgesic opioid; anti-anxiety; anti-depressant; anti-histamine; anti-hypertensive; anti-infective; anti-lipemic; anti-neoplastic; anti-platelet; anti-psychotic; anti-seizure; benzodiazepine; biologic; bronchodilator; diuretic; gastrointestinal drug; hormone; muscle relaxant; nsaid; sedative; steroid cortico; supplement. Preliminary factor analysis to determine which drugs tended to be used together yielded evidence for four factors that can be described as direct pain relievers, psychoactive-modulators, metabolic modulators, and disease treatments (e.g., infections and cancer).

To extract these drug instances two ontologies were used; RXNORM and SNOMED that were accessed through the web services at the National Library of Medicine (NLM) and the National Center for Biological Ontologies (NCBO), respectively. Using both of these it was possible to convert brand names to generic names and then extract the ontology trees from SNOMED for the generic drugs. One of the human annotators on our project had previously catalogued the MPC prescription drugs into the 23 categories given above. This was retained and the SNOMED upper level classes were mapped to these categories.

e). Comorbidities identified in progress notes

Named entity recognition of diseases and symptoms within the progress notes was undertaken to create comorbidity feature vectors (see Fig. 1). There were 61 classes in our chronic pain ontology that represented diseases and symptoms. Each document's output array was examined for words and phrases containing scores in these classes. Those, in turn, were cross-checked against the Human Disease Ontology and the Symptom Ontology from the NCBO. A total of 443,673 instances of symptoms and diseases were identified within 77,900 progress

notes. Of these, 87,287 were negated and 356,386 were affirmed. The instances represented 1,454 unique words or phrases, representing disease names or symptoms that were identified as labels or exact synonyms in either the disease or symptom ontology. These were consolidated into 545 classes represented by various child classes of these two ontologies. These can be collapsed into approximately 39 parent classes, the actual number of which depends on desired granularity. Factor analysis will be required to determine which diseases and symptoms typically cluster together. This will generate a comorbidity feature set which will part of the latent class analysis to cluster patients.

f). Clustering patients into proof-of-concept phenotype groups

Latent class analysis (mixture modeling) of a preliminary set of patient attributes (or features) has revealed interesting patient cluster characteristics. Patient attributes were obtained from progress notes and PHA questionnaires that were generated on the same patient visit. For our methods paper, we examined a use case to determine if we could extract treatments and relief that occurred in a single sentence of a progress note. The technique was highly successful, allowing us to incorporate these patient attributes in the latent class analysis together with the PHA biopsychosocial factors derived previously[1]. To simplify the presentation of the findings, we use L, M, H to represent Low, Medium, and High levels, plus some intermediated states, for the conditions or responses. The results shown below are for 9 latent classes (clusters), which is a reasonable representation of this limited set of features. It demonstrates that various unique combinations of features can be identified in reasonable fractions of the population. As we expand on these feature vectors (patient attributes) and introduce latent class analysis with structural equation models, we should be in good position to make predictions for various patient phenotypes.

When ICD-9CM codes were introduced in preliminary latent class analysis, those patients clustering in the top four treatment groups of Table 4 were about 4 to 1 more likely to suffer from non-spinal pain whereas those in the bottom 5 clusters were at least 2 to 1 more likely to suffer from spinal pain. This indicates that introducing practice management codes into the latent class analysis will be beneficial for phenotyping. Much more work on disease identification must be performed, particularly with the addition of co-morbidities that are being derived from the progress notes (see above).

Table 5. Preliminary results of latent class analysis with a subset of feature vectors consolidated into the first five columns and denoted by the top row headings.

Treatments	Depression	ADL Quality	Biopsycho-social Severity	Relief	% Population	Prediction Prob
Drug	M	M	M	M	6.5	0.93
Surgery	M	M	M	M	5.1	0.94
Drug + Surgery	L	H	L	H	4.1	0.93
Drug + Surgery Injections	H	L	H	L	11.4	0.88
Injections	L	H	L	H	12.1	0.91
Injections	M	H	M	M	14.2	0.77
Injections	L	M	M	MH	13.3	0.80
Injections	MH	ML	MH	M	20.9	0.81
Injections	M	L	MH	M	12.4	0.80

g). Patient pain reporting veracity

As with most exploratory endeavors, we have made some interesting ancillary observations. In particular, we detected pain reporting differences in sub-populations of patients. The first manuscript for this (an accepted AMIA meeting paper[4]) was submitted as an attachment in the 8th quarterly report for this grant. While we do not fully understand the determinants causing the variation in patient veracity, we anticipate that these patient differences will become part of phenotypes that we construct. It is important to note that this phenomenon would not have been detected without having both patient reports of status and outcomes together with physician reports detailed in progress notes. Furthermore, for this patient population, that was only possible because of the remarkable data repository of the Michigan Pain Consultants (MPC) and ProCare Systems, and their commitment to partner with a University to use this data for research.

In the analysis of the progress notes, pain numbers were identified in progress notes with F1 values of 0.98. Comparisons to pain reported in the PHA for the same clinical visit (Figure 5) revealed a discrepancy in some patients' veracity (nonsync group). Some patients reported pain levels to doctors (*open circles, left figure*) that were higher than those reported on the survey tool administered at the beginning of the same visit. Those patient also exhibited higher severity in other health factors (*right figure*). Therefore, the pain numbers reported to physicians were more aligned with true health status than pain values reported on the questionnaire in the waiting room. This finding was true for both sexes and provides strong support for the importance of physician recorded information, which for this important data set resides within the progress notes.

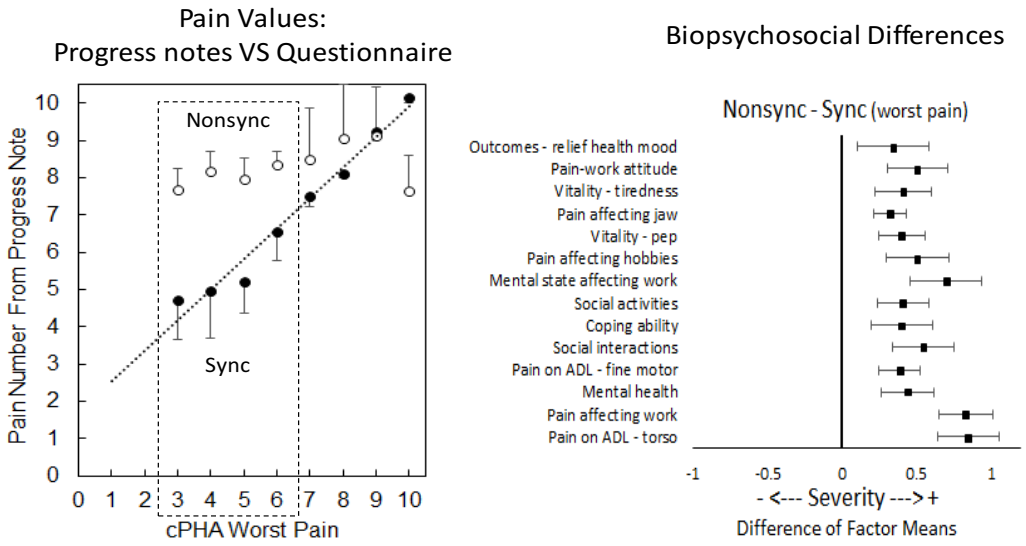

Figure 5. Differences between pain recorded in progress notes and questionnaires. (*left*) Identifying sync and nonsync populations. *right*) Biopsychosocial differences between sync and nonsync patients.

h). Longitudinal data

Evaluation of some of the characteristics of the longitudinal data available for most patients is now underway. In Table 6, we show results of content, as determined by human annotators, for a random sample of patients who had 6 to 9 progress notes on record. Important content is present within these longitudinally connected notes, particularly the results of last treatment, medications, and comments on patient behavior and affect. In Figure 6, we show that the thousands of patients with multiple progress fall into two distributions; one decreasing exponentially over time, and the other centered near 18-20 visits. A sample of the latter group reveals progress notes with similar content. The reason for this subgroup appears to be their disease severity. Since the time window for the notes is 2010-2014, that implies that more notes indicates visits occurring more often. That goes hand-in-hand with the need for more pain management. This subpopulation will be an important study cohort for severe chronic pain.

Reaching stable pain management takes time, as shown in Figure 7. This is a compilation of thousands of overlapping trajectories of patient reported relief with different start and stop points. Those that had been with the MPC practice for many years were already at steady state. Those new to the practice during our 2010-2014 study window defined the transition within the

first 1-2 years from starting pain to steady state pain. It should be noted that an improvement of 1-2 units on the 11-point Likert scale is typical for pain management services[5–7]. This also demonstrates that successful pain management takes more than one visit and more than a few months.

There are multiple types of trajectories that may exist for patients and these are likely to be strongly tied to their biopsychosocial characteristics. We anticipate that patients: can exhibit temporary relief followed by relapse; can exhibit no relief after many visits; or, can oscillate between high pain and treatment-induced low pain. These trajectories will be important to document and understand.

Table 6. Annotator analysis of 76 randomly drawn patients with 6-9 progress notes on file.

Note content	Num. of notes	% Patients
Reason for visit	296	87%
Status between visits	302	92%
Results of last treatment	370	93%
Medication usage	291	87%
Treatment given	382	92%
Other recommendations	131	71%
Comments on behavior	268	97%

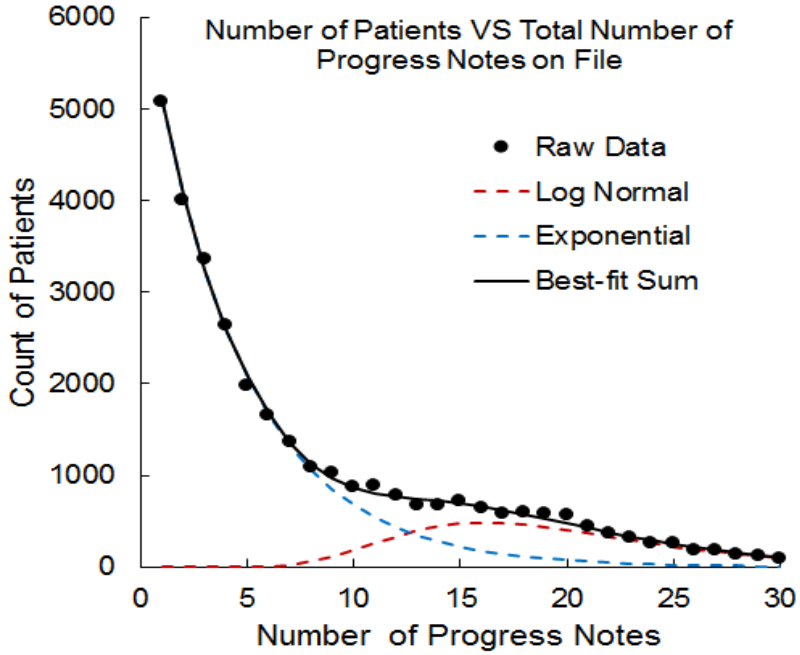

Figure 6. Patient counts versus number of notes on record. Two subpopulations exist; short-term patients, and long-term patients.

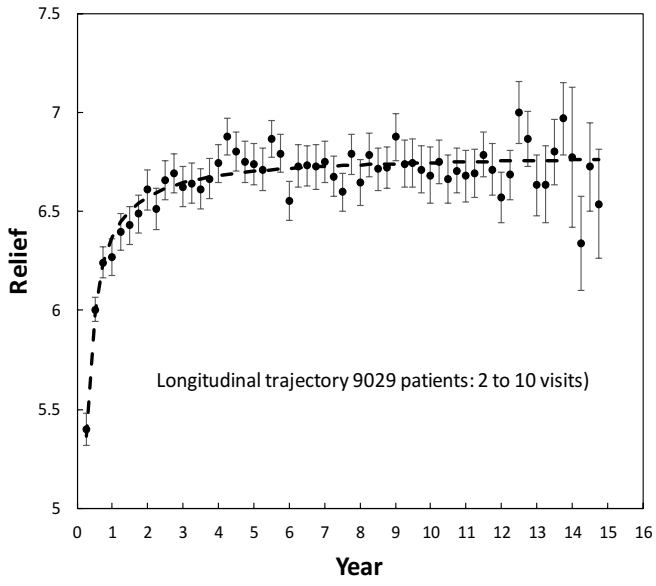

Figure 7. Compilation of longitudinal trends in patient response to the relief question in the patient-reported outcomes PHA questionnaire. The PHA was active for 4 years in the data analyzed under this grant, however many had been treated by MPC for several years. The full relief scale spans 0-10, representing 0% to 100% relief. Only a portion is shown for clarity.

i). Fuzzy Classification of Concepts Using Machine Learning

We have also been investigating the use of machine learning techniques for the classification of concepts related to pain management. It is anticipated that these methods can be used to complement the exemplar-based methods for annotation and extraction of progress note concepts important for the construction of patient phenotypes. There are several challenges that are faced when attempting to accurately classify concept phrases, regardless of methodology. We briefly describe a few challenges and the approaches we are taking in this still ongoing research.

First, concept-containing phrases can vary in length substantially and both the words and word sequence are important for phrase understanding. Therefore, it is desirable to retain both when encoding a phrase for classification. To address this, we encoded phrases as sequences of normalized difference semantic similarity indexes that represent each word as an index value, normalized to the range [-1.0,1.0], that captures the frequency with which a word appears within annotations for a specific concept. To our knowledge, this encoding technique has not been used for natural language classification tasks, although normalized difference indexes have been used for other applications[8,9]. These variable length index sequences were then used for training recurrent neural networks (RNN)[10,11] to create concept classifiers. An iterative, random sampling technique was used for RNN training with good results and reduced training time for large sets of phrases. (Space limitations prevent us from providing a full description of this methodology.)

Second, concepts are often embedded within or overlap other concept phrases. Thus, a phrase can have multiple class assignments with some likelihood. A classifier that can only

classify a phrase into one class may choose a class that is different from one assigned by a human annotator. However, when a classifier can chose alternate classes, it may be able to capture the overlapping concepts with high likelihood. To help understand the likelihood of the classifier's class assignment, RNN output was calibrated[12–14] such that it can be treated as a class membership likelihood. These likelihoods can then be used to better understand when classes overlap and when assignment to multiple classes is acceptable. As such, a more complete understanding of class membership is obtained that can be used to more accurately calculate classifier performance statistics. Moreover, the use of membership likelihoods enables a better understanding of concept overlap and relationship to other concepts that may provide useful insight for constructing patient phenotypes.

Third, classification of natural language concepts is an open set problem[15,16]. There are potentially an infinite number of concepts conveyed in written language. On the other hand, classifier training only occurs on a finite set of examples and a finite set of classes. If a phrase optimally belongs to a class not used in training then the classifier will inappropriately assign it to one of the classes used in training. These misclassifications will typically be considered false positives when computing classification statistics such as specificity or precision. We have studied this problem by removing a random set of classes during training and examining the false positive 'crosstalk' that occurs during classifier testing using all classes. We can show that likelihood thresholds can be defined to help recognize and/or remove phrases for concepts that were unknown during classifier training.

In summary, we have been investigating solutions to three challenges that impede accurate classification of concept-containing phrases. By overcoming these challenges we hope to be able to better capture the syntactic and semantic structure and relationships existing between concepts found within clinical notes and, in general, written language. As such, the classification of phrases can be better automated and enable improved processing and analysis of large repositories of clinical notes, reducing the level of manual annotation required. Furthermore, it is expected that the products of automated processing will expand our capacity to construct a more comprehensive set of patient phenotypes that can help better understand the process of managing patient pain.

2. Recap

We have endeavored to use this R21 funding to further our project to create phenotypes for people in chronic pain. Through multiple initiatives, we have made great progress extracting and uniting both the patients' and physicians' perspectives to begin understanding the biopsychosocial components that are keys to these phenotypes. This proof-of-concept, exploratory work will be critical to our future success.

We now have a viable NLP approach that should yield extensive rewards once the bulk of the progress note concepts are extracted and analyzed from the discourse-by-class arrays. We see a path forward with latent factor, latent class, and structural equation modeling, especially when applied to longitudinal data. While the longitudinal analyses and modeling cannot be done with this funding, we believe the results generated by this R21 will enable future funding to complete the modeling.

When we have successful models in hand, we will then move toward our ultimate goals of creating clinical decision support tools for the primary care setting. With the opioid catastrophe gripping medical practice, we need all the tools possible to help primary care physicians make good choices for their patients.

3. Probable additional publications generated by work done under this grant

The large methods paper that was attached to the Quarterly reports has been difficult to publish due to its complexity. This is being split into two pieces. The first will concentrate on the algorithm that generates the output array (see Figure 3 above). The relief use-case will become part of a second manuscript that expands on the concordances of the patient-reported relief outcomes and those in the progress notes. These will be linked to treatments. This will be followed with a paper on the static phenotypes generated with latent class analysis. These preliminary phenotypes will be an expansion of the work that generated Table 5 by adding more patient feature vectors from comorbidities, prescription drugs and the explicit incorporation of the ICD-9 codes. This work is well underway since the end of the no-cost extension year. We also hope to follow up on the pain reporting veracity study presented at AMIA.

4. Goals for future work made possible by this AHRQ grant

Our fundamental hypothesis is that evidence-based medicine can be, and should be, driven by data from daily practice in such fields as chronic pain medicine because the long-term associations between physician and patient are an integral part of patient response to therapy. The goal of this R21 project was to evaluate evidence relating to that hypothesis. To accomplish this required sufficiently analyzing both perspectives and assembling the results into prototype phenotypes that can provide the basis for future work. To reach our goals for the overall project, we will need to accomplish several additional steps – all made possible by the foundation created by this R21, proof-of-concept grant:

- While we have created methodologies to extract Named Entity Recognition recovery for drugs and diseases to create features for each patient, additional extractions must occur. These are all made much easier by the pre-processing of the progress notes into the word-by-class arrays.
- While we have performed latent class analysis on the factors derived from the PHA questionnaire and some of the extracted factors from the progress notes, we must finish the extraction of features and their consolidated factors to generate the most comprehensive patient clusters (phenotypes) that the data supports.
- From these phenotypes, we will need to design prototype Structural Models and evaluate them with Structural Equation Modeling approaches. This will yield early indications if Structural Models can capture the complexities of the patient phenotypes when combined with treatments and outcomes. We must also explore machine learning alternatives to the structural models. Both can generate knowledge kernels that can be used in clinical decision support engines to aid physicians in choosing the best treatments for their chronic pain patients.

List of Publications and Products

Juckett DA, Davis FN, Gostine M, Reed P, Risko R. Patient-reported outcomes in a large community-based pain medicine practice: evaluation for use in phenotype modeling. BMC Med Inform Decis Mak. 2015;15(1):41. PMID: 26017305.

Juckett D, Davis F, Gostine M, et al. Discordant patient pain level reporting between questionnaires and physician encounters of the same day. AMIA Annu Symp Proc. 2016:667-676. PMID: 28269863.

Electronic Resources

1. Juckett DA, Davis FN, Gostine M, Reed P, Risko R. Patient-reported outcomes in a large community-based pain medicine practice: evaluation for use in phenotype modeling. *BMC Med Inform Decis Mak*. 2015;15(1):41. doi:10.1186/s12911-015-0164-4.
2. Gostine M., Davis FN, Risko RJ, Gostine D, Wasan A. Outcomes Associated with Opioid Use. In: *American Association of Pain Medicine Annual Meeting*. Vol ; 2014:107.
3. Feinstein AR, Cicchetti D V. High agreement but low Kappa: I. the problems of two paradoxes. *J Clin Epidemiol*. 1990;43(6):543-549. doi:10.1016/0895-4356(90)90158-L.
4. Juckett D, Davis F, Gostine M, et al. Discordant patient pain level reporting between questionnaires and physician encounters of the same day. *AMIA Annu Symp Proc*. 2016:667-676.
5. Jensen MP, Turner J a, Romano JM, Fisher LD. Comparative reliability and validity of chronic pain intensity measures. *Pain*. 1999;83(2):157-162. http://www.ncbi.nlm.nih.gov/pubmed/10534586.
6. Tan G, Jensen MP, Thornby JI, Shanti BF. Validation of the Brief Pain Inventory for chronic nonmalignant pain. *J Pain*. 2004;5(2):133-137. doi:10.1016/j.jpain.2003.12.005.
7. Turner J a, Shortreed SM, Saunders KW, Leresche L, Berlin J a, Korff M Von. Optimizing prediction of back pain outcomes. *Pain*. 2013;154(8):1391-1401. doi:10.1016/j.pain.2013.04.029.
8. Kasten EP, Gage SH, Fox J, Joo W. The Remote Environmental Assessment Laboratory's Acoustic Library: An archive for studying soundscape ecology. *Ecol Inform*. September 2012. doi:10.1016/j.ecoinf.2012.08.001.
9. Kriegler F., Malila WA, Nalepka RF, Richardson W. Preprocessing transformations and their effects on multispectral recognition. In: *Proceedings of the 6th International Symposium on Remote Sensing of Environment*. Vol Ann Arbor, Michigan; 1969:97-131.
10. Hertz J, Krogh A, Palmer R. *Introduction to the Theory of Neural Computation*. Redwood City, CA: Addison Wesley; 1991.
11. Schellhammer I, Diederich J, Towsey M, Brugman C. Knowledge extraction and recurrent neural networks: An analysis of an Elman network trained on a natural language learning task. In: *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natuarl Language Leanrning*. Vol Sydney, Australia: Association for Computational Linguistics; 1998:73-78.
12. Niculescu-Mizil A, Caruana R. Predicting good probabilities with supervised learning. In: *Proceedings of the 22nd International Conference on Machine Learning (ICML)*. Vol Bonn, Germany; 2005:625-632.
13. Platt J. Probabilities for SV machines. In: *Advances in Large Margin Classifiers*. Vol Cambridge, Massachusetts: MIT Press; 2000:61-74.
14. Niculescu A, Caruana R. Obtaining calibrated probabilities from boosting,. In: *Proceedings of the 21st International Conference on Uncertainty in Artificial Intelligence (UAI-05)*. Vol Edinburgh, Scotland: AUAI Press; 2005:413-420.
15. Scheirer WJ, Jain LP, Boult TE. Probability models for open set recognition. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 36*. Vol ; 2014:2317-2324.
16. Scheirer WJ, de Rezende Rocha A, Sapkota A, Boult TE. Toward open set recognition. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),vol 35*. Vol ; 2013:1757-1772.