Research Paper

PAIN®

OPEN

Measuring pain care quality in the Veterans Health Administration primary care setting

Stephen L. Luther[a,b,*], Dezon K. Finch[a], Lina Bouayad[a,c], James McCart[a,d], Ling Han[e,f], Steven K. Dobscha[g,h], Melissa Skanderson[i], Samah J. Fodeh[i,f], Bridget Hahm[a], Allison Lee[e,j], Joseph L. Goulet[e,i], Cynthia A. Brandt[e,i], Robert D. Kerns[e,j]

Abstract

The lack of a reliable approach to assess quality of pain care hinders quality improvement initiatives. Rule-based natural language processing algorithms were used to extract pain care quality (PCQ) indicators from documents of Veterans Health Administration primary care providers for veterans diagnosed within the past year with musculoskeletal disorders with moderate-to-severe pain intensity across 2 time periods 2013 to 2014 (fiscal year [FY] 2013) and 2017 to 2018 (FY 2017). Patterns of documentation of PCQ indicators for 64,444 veterans and 124,408 unique visits (FY 2013) and 63,427 veterans and 146,507 visits (FY 2017) are described. The most commonly documented PCQ indicators in each cohort were presence of pain, etiology or source, and site of pain (greater than 90% of progress notes), while least commonly documented were sensation, what makes pain better or worse, and pain's impact on function (documented in fewer than 50%). A PCQ indicator score (maximum = 12) was calculated for each visit in FY 2013 (mean = 7.8, SD = 1.9) and FY 2017 (mean = 8.3, SD = 2.3) by adding one point for every indicator documented. Standardized Cronbach alpha for total PCQ scores was 0.74 in the most recent data (FY 2017). The mean PCQ indicator scores across patient characteristics and types of healthcare facilities were highly stable. Estimates of the frequency of documentation of PCQ indicators have face validity and encourage further evaluation of the reliability, validity, and utility of the measure. A reliable measure of PCQ fills an important scientific knowledge and practice gap.

Keywords: Pain, Quality, Measurement, Natural language processing

1. Introduction

Pain is a significant public health concern because of its high prevalence, inadequate pain care, and disparities.[16] Among several disadvantaged groups, military veterans have been shown to have high rates of pain and multimorbidities.[17] The Veterans Health Administration (VHA) established pain management as a high priority and established the stepped care model for pain management that prioritizes the assessment and management of most common pain conditions in the primary care setting. A 2009 VHA policy specified key dimensions for "Evaluation of Outcomes and Quality of Pain Management." After recognition of pain symptoms, key dimensions of care should be documented: timely and appropriate comprehensive pain assessment, development and enactment of a pain treatment plan, and reassessment of the effectiveness of the plan.[18,33]

Data in the electronic health record (EHR) provide an important resource for measurement of healthcare quality. Although researchers in the private sector[13,23,27] and VHA[2] have used structured (coded) data to measure quality, there are limitations, particularly in measuring the quality of care in the primary care setting.[5] One strategy shown to improve the measurement of quality in the private sector[3,15] and VHA[9,19,20] is to use natural language processing (NLP) to extract data from clinical progress notes.

Natural language processing is a range of theoretically computational techniques for analyzing and representing naturally occurring texts (either oral or written human language) at one or more levels of linguistic analysis to achieve human-like language processing for a range of tasks or applications.[24] Natural language processing began in the 1950s as the intersection of artificial intelligence and linguistics. Currently, NLP borrows from diverse fields with strategies broadly classified as top–down, typically using regular expressions or handwritten rules, or bottom–up using machine learning or statistical approaches.[26]

The NLP-based measure of pain care quality (PCQ) is informed by prior development of a reliable manual approach to identify 12 dimensions of comprehensive pain assessments, treatment, and pain reassessment.[7] The tool has been used to document

Sponsorships or competing interests that may be relevant to content are disclosed at the end of this article.

[a] Research and Development Service, James A. Haley Veterans Hospital, Tampa, FL, United States, [b] University of South Florida College of Public Health, Tampa, FL, United States, [c] Florida International University, Miami, FL, United States, [d] Morsani College of Medicine, University of South Florida, Tampa, FL, United States, [e] Pain Research, Informatics, Multimorbidities and Education Center, VA Connecticut Healthcare System, West Haven, CT, United States, [i] Department of Emergency Medicine, Yale School of Medicine, New Haven, CT, United States, [f] Department of Internal Medicine, Yale School of Medicine, New Haven, CT, United States, [g] Center to Improve Veteran Involvement in Care, VA Portland Health Care System, Portland, OR, United States, [h] Department of Psychiatry, Oregon Health & Science University, Portland, OR, United States, [j] Department of Psychiatry, Yale University School of Medicine, New Haven, CT, United States

**Corresponding author. Address: Research and Development Service, James A Haley Veterans Hospital, Mail Code 151R, 8900 Grand Oak Circle, Tampa, FL 33637-1022, United States. Tel.: 1 813 558 3966. E-mail address: steve.luther@va.gov (S.L. Luther).*

PAIN 163 (2022) e715–e724

http://dx.doi.org/10.1097/j.pain.0000000000002477

improved PCQ and relevant patient outcomes that resulted during a formative evaluation and implementation study conducted at one VHA facility.[1,25] This time intensive and costly manual approach poses a serious limitation to its scalability for performance improvement efforts. This limitation can potentially be addressed by the development of an automated NLP solution. The objective of the current study was to identify and quantify empirically derived, NLP-based PCQ indicators in the VHA. We targeted 12-specific indicators including documentation of pain, pain site, pain intensity, pain etiology or source, persistence of pain (acute or chronic), physical diagnostics, what makes pain better or worse, impact of pain on function, referral, education or self-management, and reassessment. The NLP solution was applied to data extracted from the VHA EHR in 2 time periods to provide evidence that the results were robust and could support ongoing quality improvement efforts. We also described the results across patient and facility characteristics to demonstrate stability of the results throughout the VHA. It was hypothesized that PCQ indicators could be reliably extracted from the VHA EHR using NLP.

2. Methods

2.1. Natural language processing development

2.1.1. Developing a human-annotated reference set

The study underwent ethics and regulatory reviews and approvals by each of the participating institutions' institutional review boards and research and development committees. Rule-based NLP algorithms were developed to extract the targeted PCQ indicators from progress notes written by primary care providers stored in the VHA EHR. A rule-based approach was appropriate for the task as some targets are compound concepts built on a combination of other simpler targets, and rules have good results when simple regular expressions are inadequate. The first step of the process was to develop a human-annotated (specialized chart review) set of documents (reference set) on which to train and test the NLP algorithms. To develop the reference set, we selected a sample of documents that included examples of the ways the targeted constructs are described by providers across the VHA system (**Fig. 1**). Sample documents were obtained from veterans in the VHA Musculoskeletal Disorders (MSD) Cohort. In 2013, the MSD Cohort included 5,237,763 veterans with 2 or more encounters of any ICD-9 codes for MSD between January 1, 2000, and December 31, 2014. The MSD Cohort was created to characterize variation in pain, multimorbidities, treatment, and outcomes among veterans with MSD receiving VHA care.[10] The MSD Cohort has been used to study a wide variety of pain conditions.[11,12,30] All progress notes for veterans in the MSD Cohort with outpatient visits in fiscal year (FY) 2013 (October 1, 2012-September 30, 2013) were extracted for a stratified random sample of 64 men and 13 women (women were oversampled to ensure adequate representation) from each of the 130 facilities in the VHA. To focus on clinical settings and providers targeted by VHA policy cited above, the sample was then restricted to progress notes written by providers (physicians, nurse practitioners, or physician assistants) in primary care, women's health, and geriatric primary care clinics, where the patient had a recorded pain intensity rating ≥ 4, indicative of moderate-to-severe pain intensity (n = 99,481).

From this corpus, a sample of approximately 2500 documents were selected for review in 2 waves. In the first wave, all documents from primary care visits in FY 2013 were selected for 176 patients to reflect documentation for both initial assessment of pain (eg, pain site, diagnosis, and impact on function) and pain reassessment. In the second wave, documents that contained rich information about treatments or clinician action informed by assessments (eg, patient education, medications, referrals for additional diagnostic tests, and specialty care) were selected. Examples of treatment terms were collected in the first wave and used to select documents for the second wave using the maximum number of relevant terms. The sampled notes were independently annotated by 2 clinicians, and disagreements between the 2 were adjudicated by a third subject matter expert to develop a reference set of documents for the NLP algorithm development. The annotators labeled PCQ indicators and information about the treatment of pain documented by the providers, along with contextual modifiers to help improve validity of the NLP algorithm. Examples of contextual modifiers included historical references such as "previous history of" or "while in military," hypothetical references such as "may occur" or "conditions such as" and negation such as "patient denies" or "no indication of." These efforts resulted in a total of 89,000 terms or short phrases (annotations) from which a curated vocabulary of more than 16,000 terms was developed. The curation process normalized the terms, including steps such as extracting term class pairs (original span "just standing up from chair," becomes "standing up") and cleaning extraneous textual artifacts such as punctuation and special characters. This custom vocabulary was then reviewed by subject matter experts for validity. During the process, lists of terms from standardized sources were used when possible including the VHA formulary of medications and the National Library of Medicine's Unified Medical Language System.[31]

2.1.2. Developing natural language processing algorithims

The NLP extraction algorithms were constructed in Python using the vocabulary to identify instances of the targeted PCQ indicators in the annotated documents. Notes were first preprocessed to ensure that each sentence was on a separate line. Each line was then treated as a separate unit of analysis. The text was normalized (eg, extraneous characters removed, all lower case), and a lookup was performed against the vocabulary. A sequence of annotations for each sentence was constructed, and rules were applied to the annotation sequence to identify the targeted PCQ indicators. For instance, we used the following sentence for illustration:

> *"Patient reports that gabapentin helps reduce the pain in his left knee but gets worse when he walks long distances."*

This sentence resulted in the annotation sequence: pharm><improves><pain><site><worse><aggravator>.

Based on the applied rules, the final annotations included a pharmacological treatment (gabapentin), pain mention (pain), pain site (pain in his left knee), pain reassessment (gabapentin helps reduce the pain), and an aggravator (gets worse when walking). Multiple PCQ indicators were extracted from a sentence based on how the tokens were annotated and the rules applied to the information in the sentence. For example, information about both the site and intensity of the pain might be found and extracted by separate rules from the same sentence.

The reference set was split into 2 samples with 80% (~2000 documents) being used to train the algorithms. A validation set of approximately 100 documents was held out and then used to calculate recall (sensitivity) which is true positives (true positive + false negatives), precision (positive predictive value) which is true positives (true positive + false positives), and the F measure

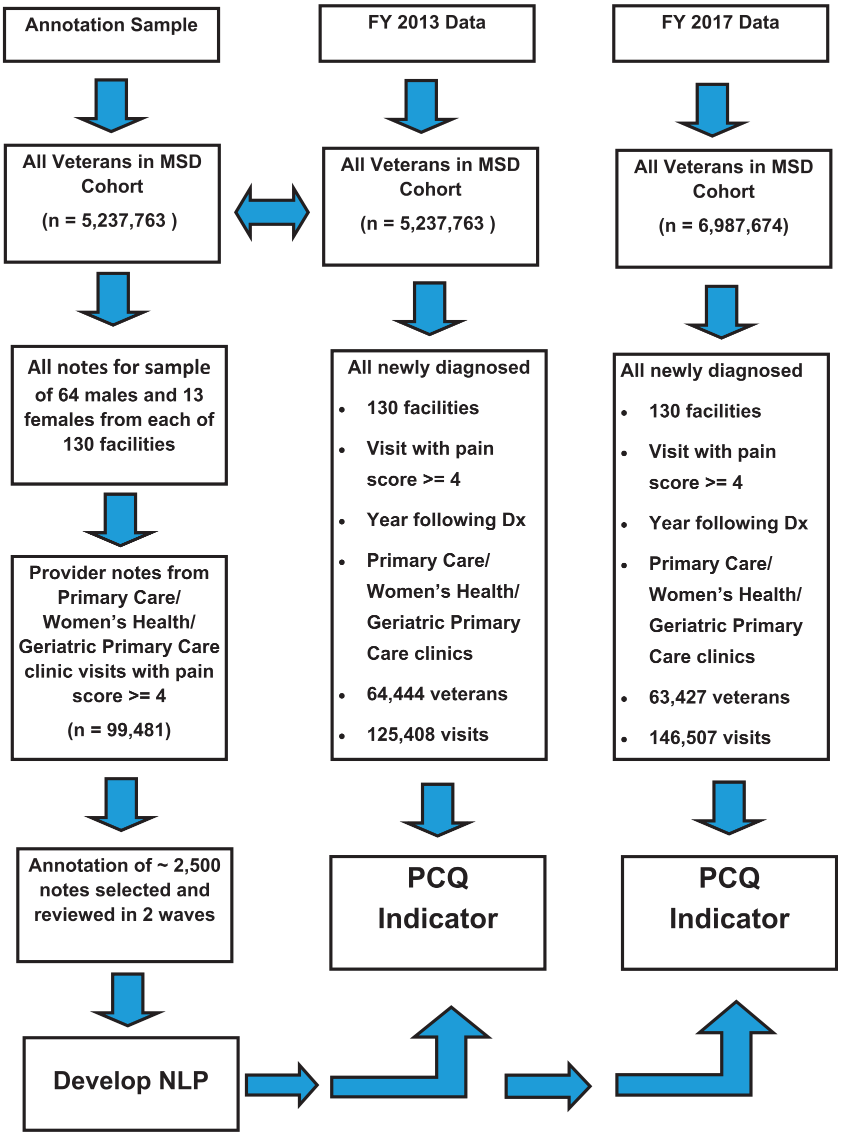
Annotation Sample
FY 2013 Data
FY 2017 Data
All Veterans in MSD Cohort (n = 5,237,763)
All Veterans in MSD Cohort (n = 5,237,763)
All Veterans in MSD Cohort (n = 6,987,674)
All notes for sample of 64 males and 13 females from each of 130 facilities
All newly diagnosed
• 130 facilities
• Visit with pain score >= 4
• Year following Dx
• Primary Care/ Women's Health/ Geriatric Primary Care clinics
• 64,444 veterans
• 125,408 visits
All newly diagnosed
• 130 facilities
• Visit with pain score >= 4
• Year following Dx
• Primary Care/ Women's Health/ Geriatric Primary Care clinics
• 63,427 veterans
• 146,507 visits
Provider notes from Primary Care/ Women's Health/ Geriatric Primary Care clinic visits with pain score >= 4 (n = 99,481)
Annotation of ~ 2,500 notes selected and reviewed in 2 waves
PCQ Indicator
PCQ Indicator
Develop NLP

Figure 1. Flow diagram describing electronic health record (EHR) data extracted to support development of natural language processing (NLP) and the subsequent national EHR samples used to demonstrate the use of the pain care quality (PCQ) measure in the Veterans Health Administration (VHA).

(precision + recall/precision × recall) between the results of the NLP algorithm and the human-annotated reference standard for individual PCQ indicators and the combined set of indicators. The F measure is commonly preferred over alternative measures of agreement such as kappa to describe the reliability of binary NLP classifications because it does not use a true negative case count that does not exist when comparing human annotation to machine annotations.[14]

2.2. Creating pain care quality indicators and total scores

Methods to calculate a PCQ score in this study were built on our previous work with a manual chart review. In that work, the PCQ was designed to assess documentation of pain assessment, treatment, and reassessment in the primary care setting.[7] A binary scoring system was used to indicate whether each indicator was present or absent in the progress notes. Provisional indicators were then presented to a multidisciplinary panel of providers and researchers who removed indicators where there was not a consensus regarding perceived importance or relevance or that overlapped with other indicators. The final chart review tool consisted of an indicator abstracting the patient's pain intensity rating at the visit and 12 dichotomously scored indicators assessing PCQ in 3 domains: assessment, treatment or plan of care, and reassessment. In the current study, we attempt to replicate this work in a national sample using NLP methods with minor refinements to the targeted PCQ indicators. Perhaps, the biggest difference relates to how documentation of treatments for pain is included as pain indicators. Because some mention of pain treatment, particularly the use of medications, is found in the "Plan of Care" section of all the documents we sampled for review, as a simple binary treatment variable, "treatment documentation yes or no" added no value to the PCQ measure. Another option we considered was to make binary indicators for specific types of treatments including medication, injections, prosthetics, and surgery. This presented 2 issues. One is the question of how to weight and subsequently sum treatment indicators coded in this manner. Is one surgery the same weight as one injection or the provision of a knee brace? Second, and perhaps more important, is the fact that it is currently not clear that the delivery of more treatments is consistent with higher quality care and improved patient outcomes. To resolve this issue, we attempted to focus on documentation about referrals that would reflect multidisciplinary pain treatment as encouraged by multiple expert panels and best practice guidance, including in VHA.[16,32,33] We targeted identification of referrals or recommendations for specialty or ancillary pain care whether or not these services were available in VHA facilities or in the community (eg, pain medicine, rehabilitation, psychological services, complementary, and integrative health approaches). We also sought to capture indicators of pain-relevant education (eg, education about pain, use of pain medications, advice regarding exercise, and explication of a biopsychosocial model of pain) and references to self-management of pain (eg, home exercise and stress reduction strategies). Finally, in addition to changes relating to treatment, we collapsed 2 closely related concepts of what makes pain better or worse into one indicator. **Figure 2** describes these changes.

As a first step in PCQ indicator measurement, a new sample of all visits for veterans from the MSD cohort (**Fig. 1**) who had their first MSD diagnosis in FY 2013 (considered incident cases) was identified (n = 107,458). From this group, veterans were eliminated if they had no primary care outpatient visits in the year after diagnosis (up to September 30, 2014), had no primary care visits with a recorded pain intensity rating ≥ 4 (consistent with moderate-to-severe pain intensity), or if there were no documents written by a primary care provider resulting in a total sample of 64,444 veterans with 125,408 outpatient visits for analysis. The final (best) NLP algorithm was applied to these data. For each visit, a binary (1 or 0) value was recorded if the PCQ indicators were documented in the progress note or not. Subsequent to the initial analyses, newer MSD Cohort data from FY 2017 (October 1, 2016-September 2017) became available, which allowed us to examine the stability of the NLP system in a separate time period. This resulted in an additional sample of 63,427 veterans and 146,507 visits, including all subsequent visits through September 30, 2018 (**Fig. 1**). The PCQ indicator extraction and analysis was repeated using FY 2017 data, and results were compared with those from FY 2013. A single document from each facility was randomly selected for review by 3 subject matter experts to validate the vocabulary used in the NLP algorithm in the newer data.

Analyses of the presence of PCQ indicators across the visits in FY 2013 and clinical judgment were used by the study team to determine the final list of PCQ indicators. The final set of PCQ indicators (n = 12) is described in **Table 1**. For descriptive purposes, the items were grouped into 3 subgroups: assessment of pain (9 indicators), plan of care (2 indicators), and reassessment (a single indicator). Also provided are examples of the text that were extracted and used to develop NLP algorithms. For each visit, a PCQ indicator's total score, (maximum = 12) and subscores (maximum = 9) described the levels and patterns of documentation.

Descriptive analysis was conducted to provide further evidence of the reliability of the NLP measure of PCQ using the FY2013 and FY2017 cohorts. Inferential statistical comparison of change over time would have required collection of longitudinal data at the physician or perhaps individual clinic level and was not our purpose here. The current study was conducted to complete the foundational work that would support such an analysis. When possible, we do, however, provide estimates of standardized effect size of differences. The percent of documentation of each PCQ indicator across all visits for each year in the study is presented along with the distribution of the total number of PCQ indicators of a possible 12 and the mean and standard deviation scores for the PCQ indicator total score. To provide evidence that the PCQ indicator total score represents a reliable summary measure, we investigated the correlation between each indicator and the total PCQ indicator score and internal consistency reliability (Cronbach alpha) for randomly selected single visits for the 64,444 veterans identified in FY 2013 and the 63,427 veterans identified in FY 2017. Descriptive analyses of PCQ indicator total scores across patient (eg, age and gender) and facility characteristics (ie, type of clinic and facility complexity) were conducted to investigate whether the documentation of PCQ indicators was consistent in patient groups across clinical settings in the VHA. The Facility Complexity Model is used for productivity and quality assurance measurement and classifies VHA facilities at levels 1a, 1b, 1c, 2, or 3 with level 1a being the most complex and level 3 being the least complex. The model has been used since 1989 and is reviewed and updated with current data every 3 years.[8] To examine reliability across the FY 2013 and FY 2017 samples, Cohen h,[4] which estimates the standardized effect size of differences in proportions taken from 2 independent samples, was calculated for all pairs of proportions for patient characteristics, patient characteristics by the number of visits, and facility complexity by visit in each year. Finally, to describe the level of variation of PCQ indicator scores across the VHA system, we calculated and rank ordered mean scores for the PCQ indicator total score and the PCQ indicator subgroups (assessment, plan of care, and reassessment) for each facility in the study. For this analysis, 7 facilities in FY 2013 and 3 facilities in FY 2017 with fewer than 120 visits in the year (on average < 10 per month) were eliminated. Statistical analyses were conducted using JMP, Version 12. SAS Institute Inc, Cary, NC.

3. Results

3.1. Natural language processing algorithm

The overall F-measure for the algorithm was established using FY 2013 data comparing human-annotated and NLP-extracted results and was found to be 91.9% with precision (positive predictive value) of 93.0% and recall (sensitivity) of 90.9%. Results for the individual PCQ indicators are presented in **Table 2**. The F-measure was ≥ 0.80 for 8 of the individual PCQ indicators with recall above 0.95 in 7 of 8. Error analysis was conducted to better understand the results among indicators with F-measures < 0.80. The intensity PCQ indicator had a recall of 0.94 and precision of 0.64. For this indicator, the NLP algorithm did a very good job identifying that intensity had been documented but specifics about levels of intensity identified by human reviewers were not as precisely identified. Error analysis suggested that the NLP algorithm often had a difficult time distinguishing between what makes pain better and self-management that affected the precision of 2 PCQ indicators including these constructs. There were a relatively large number of false positives for this indicator. The impact on function PCQ indicator had the lowest precision. Results reviewed by 3 subject matter experts found that although there were idiosyncratic differences in the vocabularies in the between the FY 2013 and FY 2017, these differences were minor, having minimal effect on the PCQ score in FY 2017.

3.2. Pain care quality indicator measurement

Documents from 130 VHA facilities, including 64,444 veterans and 124,408 unique primary care outpatient visits from FY 2013

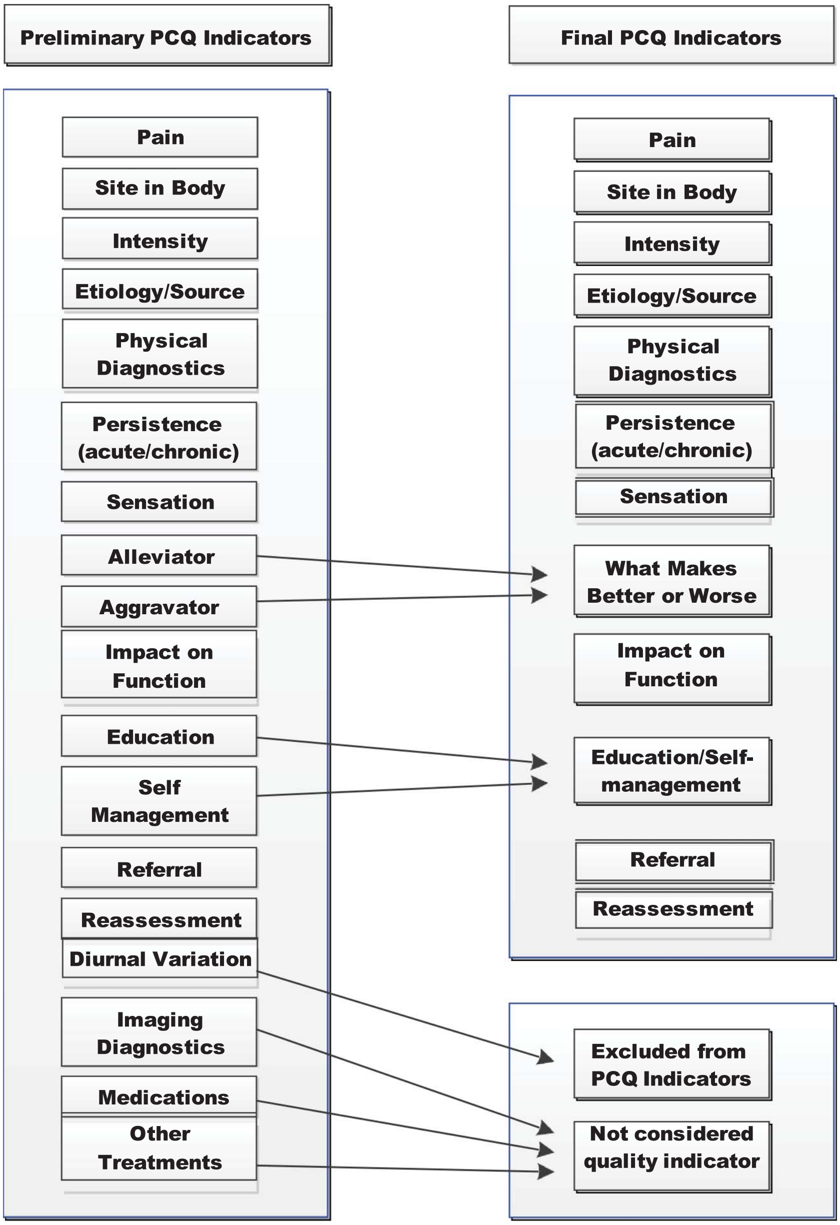

Figure 2. List of preliminarily targeted pain care quality (PCQ) indicators, those combined or dropped to create the final 12 indicators.

and 63,427 veterans and 146,507 outpatient visits from FY 2017, were analyzed. Veterans in the cohort had a mean age of 53.0 (SD = 15.5) in FY 2013 and 50.5 (SD = 16.5) in FY 2017, and they were primarily men (89.6% and 86.9%) and White (64.8% and 63.0%) with approximately half (50.3% and 50.5%) being married in each FY.

The total PCQ scores had a mean or score of 7.8 (SD = 1.9) in FY 2013 and 8.3 (2.3) in FY 2017 of a possible 12. **Table 3** describes the number of visits and percentage of documentation of each PCQ indicator for each FY. In each of the 2 years, 3 of the PCQ indicators (pain, etiology or source, and site in the body) were documented at greater than 90% of the visits. Four of the PCQ indicators (persistence, sensation, what makes pain better or worse, and pain's impact of function) were documented in fewer than half of the visits in FY 2013 and FY 2017. Standardized Cronbach alpha for the PCQ indicators was 0.61 in FY 2013 and 0.74 in FY 2017.

Detailed patient characteristics and mean results for the PCQ indicators by patient characteristics are presented in **Table 4**. The mean PCQ indicator total scores varied by no more than 0.1% across any of the demographic characteristics in either FY 2013 or FY 2017 with the exception of age, with veterans 65 or older having PCQ total scores approximately 0.5% lower than veterans younger than 65 in FY 2013. Cohen h statistic was calculated for all 30 pairs of proportions in the table. Only 2 of the comparisons, age group 50 to 64 (42.9% vs 30.3%, h = 0.22) and visits for this age group (44.6% vs 30.8%, h = 0.28), had values of h > 0.20 and < 0.50 that is defined as a "small effect" by Cohen.[4]

Table 4 also describes the proportions of patients and visits by facility complexity and type in FY 2013 and FY 2017. Of the 16 potential comparisons, only one, the numbers of women veterans being seen in the women's health clinic, (0.4% vs 10.4%, h = 0.25) met the minimum value of h = 0.20 defined as a "small effect" by Cohen. The mean PCQ total scores (**Table 4**) varied by

Table 1

Final pain care quality indicator measures.

Pain care quality indicator	Examples of the text used in NLP development
Pain assessment	
Pain	Aches and pains, are you having any pain? (yes), due to pain, flares of pain, not satisfied with her pain level, pain episodes, patients pain goal is, and reports pain
Site in body	Back of knees, bilat foot, central low back, c3-4, generalized joint, hip bursae, muscular neck, t spine, and upper thigh
Intensity	Mild, moderate, is problematic, severe, and no more than 4
Etiology or source	Bony stenosis, bulging discs, cervical pathology, djd lower leg/knee, fibromyalgia, osteoarthritic changes, and tear in lateral meniscus
Physical diagnostics	Able to feel touch, able to walk on heels and toes, Babinski normal, chronic joint deformities, decreased lumbar rom, flexed at the hip causes pain, good rom at the shoulder, and knee—tender to palpation
Persistence (eg, acute or chronic)	6 months or more, acute, almost every day, breakthrough, chronic, over 10 years, periodic, and persistent
Sensation	Shooting needles, stabbing, radiating down, dull ache, and numbness
What makes pain better or worse	Very slight movement, getting out of bed, if kneeling, in cold, pain improves as the patient gets out of bed, quiet dark room, and warm shower
Impact on function	Difficulty ambulating, discomfort when standing, hard to live a normal life, limited ROM, and sleep disturbance
Plan of care	
Referral	e-consult placed, needs to see a …, refer the patient to…, and discuss pain treatment with…
Education or self-management	Medication treatment agreement, pain education or support group, patient verbalized understanding, cold applications, daily stretching, doing lighter work, and heat or ice treatment
Reassessment	Pain improves with acupuncture and increased dose of gabapentin for pain

NLP, natural language processing.

no more that 0.1% across the 3 types of VHA clinics that provide primary care (primary care, women's health, and geriatrics care teams) in the FY 2013. In FY 2017, the mean PCQ indicator score for patients seen by geriatric care teams was approximately 1.0% lower than the other types of primary care clinics; however, this was based on a sample of less that 1000 visits.

Figure 3 describes the mean PCQ indicator score for 123 of the total 130 VHA facilities in FY 2013 and 127 of 130 facilities in FY 2017 with at least 120 outpatient visits per year sorted by mean from low to high. The mean values within each year were similar across facilities in each year. In FY 2013, the mean ranged from 6.2 (SD = 2.2) to 9.7 (SD = 1.6) with 83 (67%) falling between a mean of 7.5 and 8.5. In FY 2017, the mean ranged from 7.2 (SD = 2.8) to 9.1 (SD = 2.4) with all but 2 facilities having a mean of greater then 7.5. The pattern of documentation of PCQ indicators can be seen to vary slightly by the mean subscore across the facilities.

Table 2

Reliability of natural language processing extraction for pain care quality indicators.*

Pain care quality indicator	Precision†	Recall‡	F-measure§
Physical diagnostics	0.97	0.98	0.98
Pain	0.95	0.99	0.97
Referral	0.96	0.96	0.96
Etiology or source	0.97	0.90	0.93
Site in the body	0.79	0.98	0.87
Sensation	0.78	0.97	0.87
Persistence (eg, acute or chronic)	0.71	1.00	0.83
Reassessment	0.86	0.75	0.80
Education or self-management	0.77	0.82	0.79
Intensity	0.64	0.94	0.77
What makes pain better or worse	0.66	0.66	0.66
Impact on function	0.59	0.70	0.64

* Based on the subsample of the approximately 2500 human-annotated documents from the annotation sample taken from data from FY 2013.

† Precision = true positives or true positive + false positives.

‡ Recall = true positives or true positives + false negatives.

§ F-measure = precision + recall or precision × recall.

FY, fiscal year

4. Discussion

The lack of a reliable approach to assess quality of pain care has hindered quality improvement initiatives. We hypothesized that PCQ indicators could be reliably extracted from the VHA EHR using NLP. Our results, built on previously published work using traditional chart review methods, support this hypothesis. In this study, we estimated reliability in 2 steps, first during the NLP algorithm development and second during analysis of the PCQ indicators in a national sample of primary care visits. During the NLP development based on FY 2013 data, the reliability of the 12 PCQ indicators included in the study was very strong with an overall F-measure of 91.9%. Reliability of the individual PCQ indicators was also very good. However, for several of our PCQ indicators, both recall and precision are lower than ideal. The error analysis suggests that the lower values are likely the result of relatively rare and overlapping targets such as what makes pain better and self-management. Evolving machine-learning approaches such as deep learning may have advantages over our more traditional rule-based approach for these PCQ indicators. Deep learning, also referred to as hierarchical learning, is a machine learning method that attempts to identify data representations. Deep learning has been applied to various tasks ranging from computer vision to speech recognition and NLP.[6] Deep learning has also proven effective at accurately extracting information from the clinical text. Researchers have successfully applied deep learning to data from the EHR for tasks like information extraction, phenotyping, and outcome prediction.[29] Future research to use deep learning to improve the identification of PCQ indicators seems warranted.

Table 3

Documentation of pain care quality indicator by visit.

Pain care quality indicator	Visits		%	
	(FY 2013)	(FY 2017)	(FY 2013)	(FY 2017)
Pain	122,198	142,541	97.4	97.3
Site in the body	113,256	132,956	90.3	90.7
Etiology or source	118,068	131,148	94.1	90.0
Physical diagnostics	111,490	119,736	88.9	81.7
Intensity	81,869	110,275	65.3	75.3
Persistence (eg, acute or chronic)	52,425	73,795	41.8	50.4
Sensation (eg, pain radiates)	39,199	65,131	31.3	44.5
What makes pain better or worse	27,750	49,662	22.1	33.9
Impact on function	21,102	38,167	16.8	26.1
Referral	102,313	124,855	81.6	85.2
Education or self-management	92,733	116,203	73.9	79.3
Reassessment	99,575	114,261	79.4	78.0

FY, fiscal year.

Our PCQ indicators were developed based on VHA clinical guidelines and previous chart review at a single institution providing face validity. To provide preliminary statistical support for the use of the PCQ total score, we calculated a Cronbach alpha. For this analysis, we were able to leverage information from data in both FY 2013 and FY 2017 standardized Cronbach alpha of 0.61 was calculated for FY 2013 data and was lower than the commonly accepted target of 0.70 suggested for psychometric measures.[28] However, the standardized Cronbach alpha of 0.74 calculated from the FY 2017 data exceeded this target threshold. These results represent a more consistent documentation of PCQ indicators in FY 2017, perhaps driven by the increased emphasis on guidelines in response to the opioid crisis. Although these analyses are typically used to measure psychometrics in questionnaires and not as part of quality assurance, they are reported here to provide the reader with a frame of reference about the consistency of documentation of PCQ indicators in the VHA. Our objective is to develop a PCQ indicator measure to make it available for quality assurance and monitoring in large healthcare systems; these estimates provide real-world results based on a large sample of primary care visits for persons with moderate-to-severe musculoskeletal pain.

Although our NLP vocabulary includes freely available, open source terms from the VHA formulary and the Unified Medical Language System resources, at its core, it is built on the specialized vocabulary based on our annotation efforts. To the extent practice or documentation patterns change, the NLP vocabulary and perhaps rules would need to be updated. The results from our application to newer data from FY 2017 suggest that our current system is robust. One reason for this is our choice to de-emphasize treatments and emphasize assessment and treatment planning that is likely less likely to depend on rapidly changing technology. Even so, implementation of NLP solutions for quality assurance efforts needs to be monitored and periodically updated and refined.

We emphasized descriptive statistics in this study to provide the reader relevant information about the presence of PCQ indicators in this large cohort. Given the large sample size, relatively small differences that are likely not clinically relevant could be statistically significant.[22] We therefore refrain from presenting inferential statistics when comparing the 2 cohorts. Instead, we provide results of comparisons based on a standardized effect size measure, Cohen h. The comparison of combinations of patient, visit, and facility characteristics found only 3 of nearly 46 comparisons between FY 2013 and 2017 that met the definition of a small effect as proposed by Cohen. One of these, the increased utilization of women's health clinics, coincides with the expansion of that program in the VHA and likely reflects increased access to this clinic setting. These results are presented to support the contention that the VHA system provides a stable environment for the development and testing of our NLP PCQ measure.

The analysis of the individual and total PCQ indicator scores across patient and facility characteristics found only very small differences. This suggests that, within the context of an integrated healthcare system such as the VHA, primary care providers were found to consistently document most PCQ indicators for most patients. This lack of variability in PCQ indicator scores could alternatively be interpreted as a lack in sensitivity of the PCQ measure. However, even in the FY 2017 data, 3 of the indicators were found to be documented in fewer than 50% of the visits and therefore be targets for quality improvement efforts. In particular, impact on function was only found in 16.8% of visits in FY 2013% and 26.1% in FY 2017. Given the increasing awareness of the importance of the measurement of function in management of chronic pain, this may be an important target for organizational improvement.[21]

There are several limitations to this study that should be considered. First, although the VHA is the largest integrated healthcare system in the United States, VHA users are primarily men and share their lived experience of military service. Although our NLP-driven PCQ indicators are likely robust in non-VHA EHRs, testing and minor refinement of the systems would be necessary.

To the best of our knowledge, this study represents the first effort to automate the measure of PCQ in primary care settings in

Table 4

Patient and facility characteristics and mean pain care quality indicator total score by fiscal year.

Characteristics	N Patients (%)		N Visits (%)		PCQ score per visit mean (SD)	
	(FY 2013)	(FY 2017)	(FY 2013)	(FY 2017)	(FY 2013)	(FY 2017)
Patient						
Age						
19-34	10,949 (17.0)	14,321 (22.6)	21,125 (16.9)	31,682 (21.6)	8.0 (2.0)	8.5 (2.3)
35-49	13,375 (20.8)	15,495 (24.4)	26,882 (21.4)	35,929 (24.5)	8.0 (1.9)	8.5 (2.2)
50-64	27,617 (42.9)	19,246 (30.3) *	55,898 (44.6)	45,145 (30.8) *	7.8 (1.9)	8.2 (2.4)
65-79	9924 (15.4)	11,994 (18.9)	17,236 (13.7)	26,206 (17.9)	7.6 (1.9)	8.0 (2.3)
80+	2579 (4.0)	2371 (3.7)	4267 (3.4)	4836 (3.3)	7.4 (1.9)	8.3 (2.3)
Gender						
Female	6692 (10.4)	8504 (13.4)	14,182 (11.3)	21,046 (14.4)	7.8 (1.9)	8.4 (2.3)
Male	57,752 (89.6)	54,923 (86.6)	111,226 (88.7)	125,461 (85.6)	7.9 (1.9)	8.4 (2.3)
Race						
Non-White	12,594 (35.2)	23,553 (37.1)	43,855 (35.0)	55,694 (38.0)	7.8 (1.9)	8.3 (2.3)
White	41,761 (64.8)	39,874 (62.9)	81,553 (65.0)	90,813 (62.0)	7.9 (1.9)	8.3 (2.3)
Current marital status†						
Married	32,438 (50.3)	32,029 (50.5)	61,695 (49.0)	83,944 (49.8)	7.9 (1.9)	8.3 (2.3)
Unmarried	32,006 (49.7)	31,388 (49.5)	61,713 (51.0)	84.457 (50.01)	7.8 (1.9)	8.3 (2.4)
Current smoking†						
Smoker	27,887 (43.3)	25.573 (42.4)	56,451 (45.0)	59,262 (40.4)	7.9 (1.9)	8.3 (2.3)
Nonsmoker	35,667 (55.4)	34,852 (57.6)	63,636 (50.7)	80,167 (54.7)	7.8 (1.9)	8.3 (2.3)
Obesity†						
Obese‡	29,017 (45.0)	29,612 (44.1)	57,345 (45.7)	70,163 (47.9)	7.9 (1.9)	8.4 (2.3)
Not obese	35,541 (55.1)	33,815 (53.3)	66,664 (53.2)	72,677 (49.6)	7.8 (1.9)	8.3 (2.4)
Facility§						
Facility complexity‖						
1a	28,749 (44.6)	28,793 (45.4)	56,725 (45.2)	65,877 (47.1)	7.9 (1.9)	8.3 (2.3)
1b	13,278 (20.6)	14,237 (22.0)	24,893 (19.9)	24,881 (17.8)	7.8 (1.9)	8.4 (2.4)
1c	8809 (13.6)	8168 (12.8)	16,139 (13.3)	19,709 (14.1)	7.6 (2.0)	8.4 (2.3)
2	7557 (11.7)	7014 (11.1)	15,128 (12.1)	17,127 (12.2)	7.6 (1.9)	8.4 (2.4)
3	5959 (9.2)	5200 (8.2)	11,784 (9.4)	12,263 (8.8)	7.6 (1.9)	8.2 (2.4)
Primary stop code						
Primary care or medicine¶	62,213 (96.5)	62,193 (98.0)	119,997 (95.6)	138,507 (94.5)	7.8 (1.9)	8.3 (2.3)
Women's health¶	2601 (0.4)	6580 (10.4)*	4816 (3.9)	7043 (4.8)	7.9 (1.9)	8.5 (2.4)
Geriatric care teams¶	346 (0.05)	1.5	615 (0.5)	957(0.6)	7.8. (2.2)	7.3 (2.7)

* Cohen h > .20 and < .50, defined as small effect.

† Unknown values for race, marital status, current smoker, and obesity ranged from 1.1% to 4.5% across the years studied.

‡ Obesity = BMI ≥ 30.

§ Facilities, n= 130 2.Visits, n = 124,408 (FY 2013), n -= 146,507 (FY 2017).

‖ The VHA Facility Complexity Model was first adopted for use in 1989. The Facility Complexity Model classifies VHA facilities at levels 1a, 1b, 1c, 2, or 3 with level 1a being the most complex and level 3 being the least complex. The model is reviewed and updated with current data every 3 years. http://opes.vssc.med.va.gov/Pages/Facility-Complexity-Model.aspx.

¶ Not mutually exclusive between stop codes.

FY, fiscal year; PCQ, pain care quality; VHA, Veterans Health Administration.

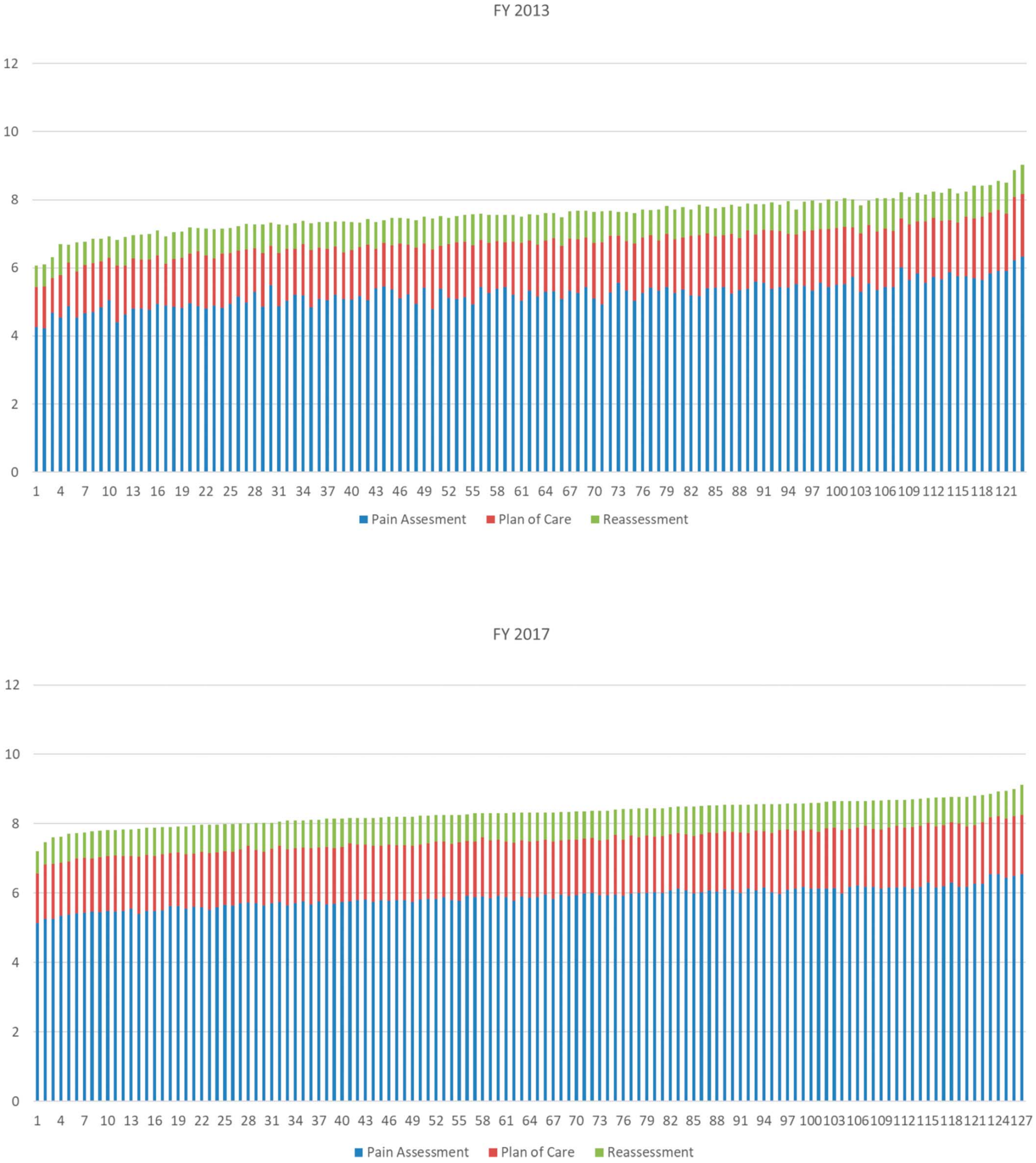

Figure 3. Mean pain care quality (PCQ subscore (Y axis) by individual Veterans Health Administration (VHA) facility (X axis) by fiscal year (FY) 2013 and 2017. The mean contributions of each PCQ indicator subscore are represented by colored bars (blue = pain assessment subscore, red = plan of care subscore, and green = reassessment subscore).

a large integrated healthcare system. We report relatively simple analyses on a large cohort of patients with clinically significant pain. Further refinement and evolution of the PCQ indicators are likely warranted. This work should include multivariable statistical analyses, perhaps including information about individual providers, which is beyond the scope of the current project.

5. Conclusion

Combining evidence of documentation of individual PCQ indicators extracted from progress notes into a total score at the visit level may provide healthcare systems a potentially useful tool for monitoring and improving quality of care. The consistency of a visit-level score based on results of our NLP suggests that this is a valid approach. In this study, we include the PCQ indicator for reassessment in the visit-level score although distinguishing reassessment from an initial assessment is not possible unless assessment has occurred in the previous visit. We did this because although our sampling frame attempted to identify more recently diagnosed patients, we could not determine the exact first visit and thereby the episode of care for specific MSD diagnoses. If measures of PCQ indicators are implemented in ongoing systems to monitor care, efforts should be made to do so and adjust the PCQ indicator score accordingly.

Conflict of interest statement

The authors have no conflicts of interest to declare.

Acknowledgements

This work was supported by grants awarded by the National Center for Complementary and Integrative Health (NCCIH) at the National Institutes of Health, Health Services Research & Development Service (HSR&D) at the Department of Veterans Affairs, and Veteran Health Administration (VHA) (NCCIH R01 AT008448/HSR&D IIR 14-438; Brandt, Kerns, Luther PIs, and HSR&D CRE 12-012; Goulet, Brandt PIs). The views expressed here are those of the authors and do not necessarily represent the official policy or position of the Department of Veterans Affairs, NCCIH, or the U.S. Government.

Article history:
Received 30 June 2020
Received in revised form 5 August 2021
Accepted 18 August 2021
Available online 15 September 2021

References

[1] Anderson DR, Zlateva I, Coman EN, Khatri K, Tian T, Kerns RD. Improving pain care through implementation of the Stepped Care Model at a multisite community health center. J Pain Res 2016;9:1021–9.

[2] Bravata DM, Myers LJ, Cheng E, Reeves M, Baye F, Yu Z, Damush T, Miech EJ, Sico J, Phipps M. Development and validation of electronic quality measures to assess care for patients with transient ischemic attack and minor ischemic stroke. Circ Cardiovasc Qual Outcomes 2017;10:e003157.

[3] Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, Bekhuis T, Jacobson RS. Automated annotation and classification of BI-RADS assessment from radiology reports. J Biomed Inform 2017;69:177–87.

[4] Cohen J. Statstical power analyses for the behavioral sciences (second addition). Hillsdale: Lawrence Erlbaum Associates, 1988.

[5] Cohen DJ, Dorr DA, Knierim K, DuBard CA, Hemler JR, Hall JD, Marino M, Solberg LI, McConnell KJ, Nichols LM. Primary care practices' abilities and challenges in using electronic health record data for quality improvement. Health Aff 2018;37:635–43.

[6] Deng L, Yu D. Deep learning: methods and applications. SIG 2014;7: 197–387.

[7] 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. Transl Behav Med 2014;4:184–9.

[8] Fact Sheet, Facility Complexity Model. VHA Office of Productivity, Efficiency and Staffing (OPES). Available at: http://opes.vssc.med.va.gov/Pages/Facility-Complexity-Model.aspx. Accessed March 14, 2020.

[9] Garvin JH, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, Heidenreich P, Bolton D, Heavirland J, Kelly N. Automating quality measures for heart failure using natural language processing: a descriptive study in the department of veterans affairs. JMIR Med Inform 2018;6:e5.

[10] Goulet JL, Kerns RD, Bair M, Becker W, Brennan P, Burgess DJ, Carroll C, Dobscha S, Driscoll M, Fenton BT. The musculoskeletal diagnosis cohort: examining pain and pain care among veterans. PAIN 2016;157: 1696.

[11] Hausmann LR, Brandt CA, Carroll CM, Fenton BT, Ibrahim SA, Becker WC, Burgess DJ, Wandner LD, Bair MJ, Goulet JL. Racial and ethnic differences in total knee arthroplasty in the Veterans Affairs health care system, 2001–2013. Arthritis Care Res 2017;69:1171–8.

[12] Higgins DM, Kerns RD, Brandt CA, Haskell SG, Bathulapalli H, Gilliam W, Goulet JL. Persistent pain and comorbidity among operation enduring freedom/operation Iraqi freedom/operation new dawn veterans. Pain Med 2014;15:782–90.

[13] Hirsch AG, Scheck McAlearney A. Measuring diabetes care performance using electronic health record data: the impact of diabetes definitions on performance measure outcomes. Am J Med Qual 2014;29:292–9.

[14] Hripsak G, Rothschilds AS. Agreement, the F-measure, and reliability in information retrieval. J Am Med Inform Assoc 2005;12:296–8.

[15] Imler TD, Sherman S, Imperiale TF, Xu H, Ouyang F, Beesley C, Hilton C, Coté GA. Provider-specific quality measurement for ERCP using natural language processing. Gastrointest Endosc 2018;87:164–73.

[16] Institute of Medicine (US) Committee on Advancing Pain Research, Care, and Education. Relieving Pain in America: A Blueprint for Transforming Prevention, Care, Education, and Research. Washington, DC: National Academies Press (US), 2011. PMID: 22553896.

[17] Kerns RD, Otis J, Rosenberg R, Reid MC. Veterans' reports of pain and associations with ratings of health, health-risk behaviors, affective distress, and use of the healthcare. J Rehabil Res Dev 2003;40:371–80.

[18] Kerns RD, Philip EJ, Lee AW, Rosenberger PH. Implementation of the veterans health administration national pain management strategy. Transl Behav Med 2011;1:635–43.

[19] Kerr GS, Richards JS, Nunziato CA, Patterson OV, DuVall SL, Aujero M, Maron D, Amdur R. Measuring physician adherence with gout quality indicators: a role for natural language processing. Arthritis Care Res 2015; 67:273–9.

[20] Kim Y, Garvin JH, Goldstein MK, Hwang TS, Redd A, Bolton D, Heidenreich PA, Meystre SM. Extraction of left ventricular ejection fraction information from various types of clinical reports. J Biomed Inform 2017; 67:42–8.

[21] Kroenke K, Krebs EE, Turk D, Von Korff M, Bair MJ, Allen KD, Sandbrink F, Cheville AL, DeBar L, Lorenz KA, Kerns RD. Core outcome measures for chronic musculoskeletal pain research: recommendations from a Veterans Health Administration work group. Pain Med 2019;20:1500–8.

[22] Lantz B. The large sample size fallacy. Scand J Caring Sci 2013;27: 487–92.

[23] Laws MB, Michaud J, Shield R, McQuade W, Wilson IB. Comparison of electronic health record–based and claims-based diabetes are quality measures: causes of discrepancies. Health Serv Res 2018;53:2988–3006.

[24] Liddy ED. Natural language processing in Encyclopedia of Library and Information Science 2nd ed., New York: Marcel Decker, 2001.

[25] 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 Rehabil Res Develop 2016;53:137–46.

[26] Nadkarni PM, Ohno-Machado L, Chapman WW. Natural language processing: an introduction. J Am Med Inform Assoc 2011;18:544–51.

[27] Newman ED, Lerch V, Billet J, Berger A, Kirchner HL. Improving the quality of care of patients with rheumatic disease using patient-centric electronic redesign software. Arthritis Care Res 2015;67:546–53.

[28] Nunnally JC, Bernstein IR. Psychometric theory. New York: McGraw-Hill, 1994.

[29] Shickel B, Tighe PJ, Bihorac A, Rashidi P, Deep EHR. A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. IEEE J Biomed Health Inform 2018;22:1589–604.

[30] Sinnott PL, Dally SK, Trafton J, Goulet JL, Wagner TH. Trends in diagnosis of painful neck and back conditions, 2002 to 2011. Medicine 2017;96:e6691.

[31] The Unified Medical Language System (UMLS). Available at: https://www.nlm.nih.gov/research/umls/new_users/online_learning/OVR_001.html. Accessed May 12, 2021.

[32] U.S. Department of Health and Human Services. Pain management best practices inter-agency task force report, 2019. Available at: https://www.hhs.gov/sites/default/files/pmtf-final-report-2019-05-23.pdf. Accessed May 10, 2021.

[33] Veterans Health Administration. VA pain management directive (2009-053). Washington: Department of Veterans Affairs, 2009.