Journal of Biomedical Informatics 120 (2021) 103864

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Original research

Development of a generalizable natural language processing pipeline to extract physician-reported pain from clinical reports: Generated using publicly-available datasets and tested on institutional clinical reports for cancer patients with bone metastases

Hossein Naseri [a,b,*], Kamran Kafi [c], Sonia Skamene [d], Marwan Tolba [d], Mame Daro Faye [d], Paul Ramia [d], Julia Khriguian [d], John Kildea [a,b]

[a] Medical Physics Unit, McGill University Health Centre, Montreal, QC, Canada
[b] Physics Department, McGill University, Montreal, QC, Canada
[c] Imagia, 6650 Saint-Urbain Street, Suite 100, Montreal, QC, Canada
[d] Radiation Oncology, McGill University Health Centre, Montreal, QC, Canada

ARTICLE INFO

Keywords:
Natural Language Processing
Generalizabile
MetaMap
Pain
Bone metastasis

ABSTRACT

Objective: The majority of cancer patients suffer from severe pain at the advanced stage of their illness. In most cases, cancer pain is underestimated by clinical staff and is not properly managed until it reaches a critical stage. Therefore, detecting and addressing cancer pain early can potentially improve the quality of life of cancer patients.

The objective of this research project was to develop a generalizable Natural Language Processing (NLP) pipeline to find and classify physician-reported pain in the radiation oncology consultation notes of cancer patients with bone metastases.

Materials and Methods: The texts of 1249 publicly-available hospital discharge notes in the i2b2 database were used as a training and validation set. The MetaMap and NegEx algorithms were implemented for medical terms extraction. Sets of NLP rules were developed to score pain terms in each note. By averaging pain scores, each note was assigned to one of the three verbally-declared pain (VDP) labels, including no pain, pain, and no mention of pain. Without further training, the generalizability of our pipeline in scoring individual pain terms was tested independently using 30 hospital discharge notes from the MIMIC-III database and 30 consultation notes of cancer patients with bone metastasis from our institution's radiation oncology electronic health record. Finally, 150 notes from our institution were used to assess the pipeline's performance at assigning VDP.

Results: Our NLP pipeline successfully detected and quantified pain in the i2b2 summary notes with 93% overall precision and 92% overall recall. Testing on the MIMIC-III database achieved precision and recall of 91% and 86% respectively. The pipeline successfully detected pain with 89% precision and 82% recall on our institutional radiation oncology corpus. Finally, our pipeline assigned a VDP to each note in our institutional corpus with 84% and 82% precision and recall, respectively.

Conclusion: Our NLP pipeline enables the detection and classification of physician-reported pain in our radiation oncology corpus. This portable and ready-to-use pipeline can be used to automatically extract and classify physician-reported pain from clinical notes where the pain is not otherwise documented through structured data entry.

1. Introduction

Two-thirds of cancer patients with advanced metastatic disease experience pain [1], and nearly 50% of these patients identify pain as a significant problem that deteriorates their quality of life [2,3]. Pain can also induce stress that may suppress the immune system. For instance, it has been demonstrated that pain in metastatic patients can suppress the natural killer cells that control tumor growth and metastasis [4]. Because of these issues, several organizations such as the World Health Organization (WHO) and the American Pain Society recommend that

* Corresponding author at: Medical Physics Unit, McGill University Health Centre, Montreal, QC, Canada.
E-mail address: hossein.naseri@mail.mcgill.ca (H. Naseri).

https://doi.org/10.1016/j.jbi.2021.103864
Received 27 November 2020; Received in revised form 30 June 2021; Accepted 4 July 2021
Available online 12 July 2021

physicians properly document pain in Electronic Health Records (EHRs) to facilitate best practice pain management, follow up, and quality assurance [1,5–7].

Consultation notes in EHRs represent a wealth of useful information on patients' health and outcomes. But, due to their largely unstructured nature and typically non-standardized formatting, extracting useful information from these unstructured free-text documents efficiently, is a challenging task [8]. This may result in consultation notes being ignored or not optimally used in clinical cancer management and outcomes research.

One potential approach to meet this challenge is to adopt Natural Language Processing (NLP) pipeline to parse consultation notes. This approach is the subject of our presently-reported study with a focus on pain mentions.

1.1. NLP for pain assessment

NLP is a branch of Computer Science that utilizes statistical functions and computational algorithms to analyze unstructured free text and extract quantitative information from it [9]. Algorithms can be trained to process large corpora of clinical narratives and extract relevant biomedical information from them. To extract biomedical concepts from clinical texts, one approach is to use pre-trained NLP models such as bidirectional encoder representations from transformers (BERT) [10]. Another approach is to combine the NLP technique with structured databases of clinical terminologies. Such structured databases are designed to categorize and classify medical terms and clinical information into standardized tables with a unique code for each medical concept.

There are several well-known databases of clinical terminologies in-use worldwide. The International Classification of Diseases (ICD) [11] is one of them, maintained by the WHO. ICD-11 is the latest available update of the ICD database. The SNOMED CT is the next one that has encoded over 340,000 multilingual clinical healthcare terminologies [12]. This database is maintained by the SNOMED International association. The Unified Medical Language System (UMLS) [13] is another database maintained by the US National Library of Medicine (NLM). The UMLS provides standard codes for thousands of biomedical concepts and it includes both the ICD and SNOMED CT vocabularies [14]. The NLM also provides the MetaMap NLP tool [15,16] to extract biomedical concepts from clinical notes and map them to the UMLS database. MetaMap, which is widely utilized in medical NLP applications [17,18], has built-in libraries for sentence segmentation, concept tokenization and abbreviation/acronym identification [19]. MetaMap uses the NegEx [20] negation detection algorithm to determine whether mentions of medical terms in the corpus were negated. NegEx has a superior performance in negation detection compared to other algorithms [21].

NLP techniques have been used for medical keyword searches, classification of diagnoses, and extraction of cancer phenotype and symptom-related information from clinical notes [22–28]. In some studies, NLP has been used to extract mentions of chest pain and back pain [29,30]. NLP has also been deployed to identify and classify chronic pain [31,32], and to extract cancer-related pain scores [33]. Eisman et al.[34] successfully implemented the pre-trained BERT model to extract angina symptoms from patients' clinical notes. Bui and Zeng [35] developed an NLP algorithm using regular expression analyzes to extract pain terminologies from clinical texts. Then, the authors classified each note into "pain" and "no pain" groups using supervised machine learning method. However, their algorithm was limited to explicit indications of "pain" and did not achieve accuracy higher than 79% in identifying and assigning pain scores. Heintzelman et al. [33] developed a more robust rules-based NLP technique to process clinical notes and detect all pain terms and their severity scores in each note in their cancer dataset with an accuracy of 96%. Then, for each note they considered the pain term with the maximum severity as the "pain index" and used it to evaluate the correlation between the cancer pain severity and survival rate in metastatic prostate cancer patients. However, upon testing on a publicly-available hospital discharge summary corpus, the accuracy of their NLP algorithm dropped to 64%. Also, the authors of Ref. [33] found that their algorithm needed to be trained on the new pain description patterns that they found in the publicly-available corpora. The authors argued that this lack of generalizability was attributed to more complex hypothetical wordings and past tense descriptions in publicly-available corpora compared to cancer data sets. It has been shown that more generalizable text classification models can be achieved by exploiting word embedding techniques [36,37]. In study by Tao et al. [38], integration of the GloVe word embedding resulted in a significant performance improvement in the generalizability of extracting prescription information (medication names, dosages and frequencies) from clinical notes. Testing on the i2b2 dataset, authors showed that F-1 score of their algorithm increased from 0.78 to 0.83 when they integrated GloVe word embedding.

The objective of our study was to develop a generalizable (i.e. dataset independent) NLP pipeline to retrospectively process patients' medical notes and identify all pain terms and their severity scores in each note and assign a single verbally-declared pain (VDP) to each note, representing the overall pain of the patient at the time of the consultation. For each note, our VDP was obtained by averaging over the pain scores detected in the note. For generalizability, unlike Heintzelman et al. [33], we first trained our pipeline on a publicly-available dataset, and afterward applied our trained pipeline on another publicly-available dataset and on our institutional radiation oncology dataset. Moreover, motivated by the findings of Tao et al. [38], and in order to provide a more generalizable solution, we used distributed word vectorization methods and word similarity features (GloVe word embedding). Also, unlike Heintzelman et al. [33], that used pain term with the highest pain-score as their pain index, we averaged the pain scores to assign a VDP to each note. We showed that these methods enabled building a database-independent pipeline to identify pain description patterns, exclude irrelevant mentions of pain, and calculate the physician-reported VDP at the time of the hospital visit. This is important as it now allows pain to be reliably extracted from radiation oncology consultation notes in a way that can facilitate further pain-related studies.

The pain-related terms used in this paper are defined in Table 1.

2. Materials and methods

2.1. Corpora

In this study we used three independent corpora to develop and test our NLP pipeline: (i) 1249 discharge summaries from the Informatics for Integrating Biology & the Bedside (i2b2) #1 A Smoking challenge database [39,40], (ii) 30 discharge summaries from the Medical Information Mart for Intensive Care III (MIMIC-III) database [41], and (iii) 788 consultation notes from the EHRs of 462 metastatic cancer patients previously treated at our institution. Consultation notes for metastatic cancer patients from our institution were extracted from the ARIA database for Radiation Oncology (Varian Medical Systems, Palo Alto, CA). All patients in our institutional corpus received palliative radiotherapy for a secondary malignant neoplasm of bone at our cancer center between January 2016 and September 2019. The textual data from our institutional corpus were extracted from Microsoft Word (.doc) documents using the Python textract package [42].

Detailed descriptions of the three corpora are presented in Appendix 7.1, and details of the number of characters and words in each corpus are presented in Table 15.

All three corpora had similar mean numbers of characters per clinical note (between 7000 to 9000, which is equivalent to two or three pages of single-spaced text).

As presented in Fig. 2, of the 1249 i2b2 notes, 1099 randomly-selected notes were used for concept extraction and training the NLP

Table 1
Definitions of the terms used in this paper.

Term	Definition
Pain terms	The pain-related medical terms that were collected in Table 2. Each note might contain multiple pain terms.
Pain concepts	The UMLS medical concepts that which were obtained by mapping the pain terms to the UMLS metathesaurus (Table 19). Multiple pain concepts might be mapped to one pain term.
Pain score	A pain term in a phrase that explicitly indicates an experience (score 1) or denial (score 0) of pain at the time of the hospital visit. Pain terms that were not related to the time of the visit were assigned as irrelevant pain. (See Fig. 4) Each note might contain multiple pain scores.
VDP	A three-point verbally-declared pain (VDP) (no mention of pain, pain, no pain) that was assigned to each note by averaging valid pain scores. (See Section 2.3.3)

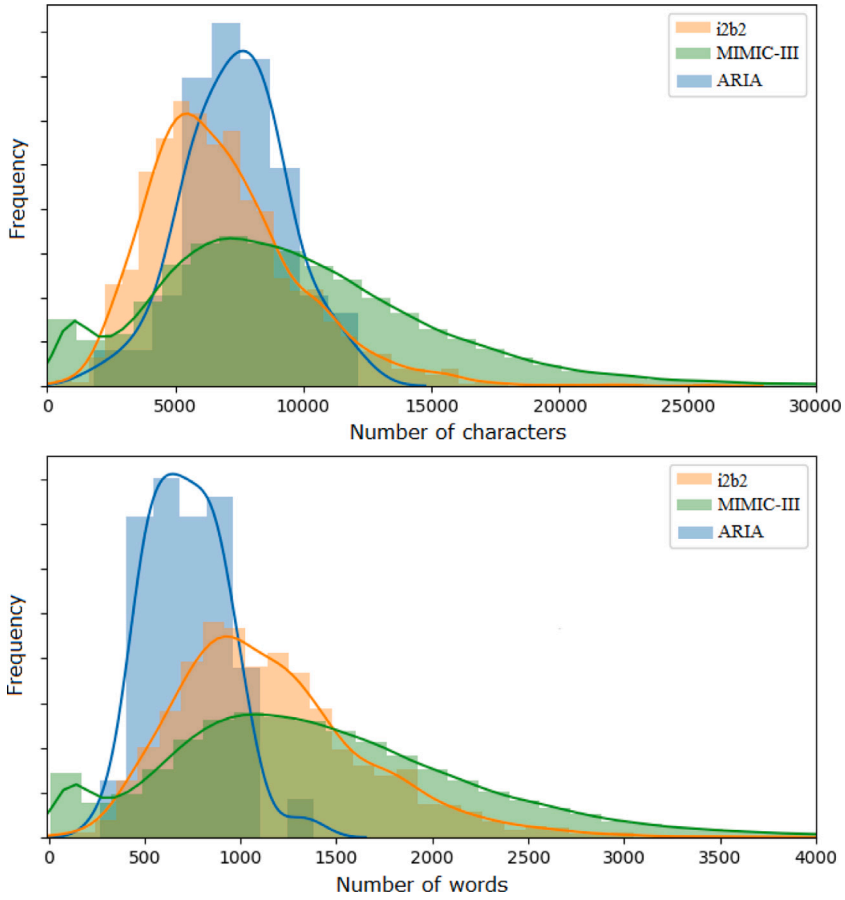

Fig. 1. Normalized distributions of the number of characters (top panel) and number of the words (bottom panel) in the i2b2 (shown by orange color), MIMIC-III (green), and ARIA (blue) corpora. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

pipeline, 120 notes (4 sets of 30 notes) were used for validation, and the remaining 30 randomly-selected notes were reserved for testing. In each iteration of training, we did a performance evaluation on one set (30 notes) from the validation corpus. The test corpus was used for final performance evaluation once the pipeline was completely developed. Later, 30 notes from the MIMIC-III and 30 notes from ARIA corpora were used for testing of the generalizability of the fully-developed NLP pipeline. It should be noted that the MIMIC-III and ARIA corpora were not used in any of the iterations of the training and validation. Another set of 150 notes from ARIA corpora were used for testing the performance of our NLP pipeline in assigning a Verbally-declared pain (VDP) label to each note. Cochran's [43] sample size formula was used to determine the confidence interval of the selected sample sizes, as presented in Section 7.4, in the Supplementary Information.

2.2. Preparation of the validation and test corpora

The notes from the validation corpus were annotated by developers and were used to evaluate the performance of our NLP pipeline in four iterations of the training. The final performance of the pipeline to detect and score pain was evaluated against an expert-annotated (gold-standard) test corpus from each dataset (Fig. 2).

We extracted all the sentences from each set in the validation and test corpus. The sentences from validation set 1, set 2 and set 3 (2310, 2332, 2075 sentences, respectively) were manually annotated by the primary developer. Validation set 4 (1012 sentences) was manually annotated by an independent developer. The sentences from the i2b2 (2361 sentences) and MIMIC-III (2717 sentences) test corpora were manually annotated by an MD physician. The sentences from the ARIA test corpus (1132 sentences) were manually annotated by a radiation oncologist at our institution. The selected sample size resulted in 95% confidence level with less than 1% margin of error (the sample size calculation is presented in Section 7.4 in the Supplementary Information).

Following Heintzelman et al.'s [33] example, sentences from the test sets were annotated by our NLP pipeline. The domain experts (MD physician, radiation oncologist) were then asked to compare their

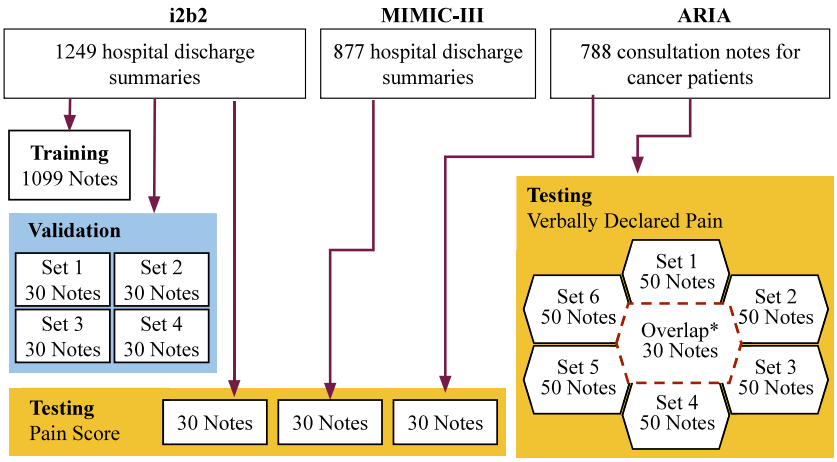

Fig. 2. The corpora used in this paper. From 1249 hospital discharge summaries form i2b2 corpora, 1099 notes were used for concept extraction and training of our NLP pipeline, 120 annotated notes were used for validation of our NLP pipeline in four iterations and 30 notes were reserved for testing of the fully-developed pipeline. The MIMIC-III and ARIA corpora were used only for testing of our NLP pipeline (these two corpora were never used for training). 150 notes from the ARIA corpora were used for testing the Verbally-declared pain (VDP) classification method.

manually-annotated sentences against the NLP annotation results to produce the gold-standard test sets. The rational for this step was to ensure that the experts did not accidentally miss or mislabel any pain term.

To evaluate the accuracy of our VDP classification method, another independent set of 150 notes from the ARIA corpus was annotated by six annotators (one oncologist, one medical physicist, and four oncology residents). Each annotator was asked to annotate a set of 50 notes consisting of 20 unique notes and 30 notes that were shared among all six annotators. These 30 notes were used to report inter-annotator agreement using Fleiss' kappa statistical measure [44]. Each annotator was asked to review each note and assign it to one of the five-grade verbal rating scales; no mention of pain (when pain was not reported in the note or pain was not reflecting the current state of the illness), no pain (if the pain was explicitly denied), mild (pain score 1–3), moderate (pain score 4–6) and severe (pain score 7–10). However, since we found that pain scores were not consistently documented in the radiation oncology consultation notes, which led to poor kappa measures for inter-annotator agreement, we instead defined a three-grade verbally-declared pain scale (VDP) incorporating 'no mention of pain', 'no pain', and 'pain' (by grouping mild, moderate and severe pain scales as 'pain'). The 150 VDP-annotated notes provided a gold-standard for evaluation of the accuracy of our VDP classification method. The selected sample size resulted in 0.026 standard error within a 95% confidence interval. The detailed sample size calculation can be found in Section 7.4 of the Supplementary Information.

Because the aim of this project was classifying cancer pain in radiation oncology clinical notes, the accuracy of our VDP classification method was only tested on the ARIA corpus. Given the effort required, we did not ask the radiation oncologists to spend their time annotating i2b2 and MIMIC-III corpora.

2.3. Pain detection pipeline

Our pain detection pipeline consisted of three parts: (1) an NLP pipeline to extract all UMLS medical concepts from the text documents, (2) a rules-based classifier to identify pain terms and extract valid pain scores, and (3) a method to calculate an average pain intensity and assign a physician-reported VDP to each note. The terms used in this paper are defined in Table 1.

2.3.1. Step 1: UMLS medical concept extraction

A flowchart describing our medical concept extraction pipeline, is provided in Fig. 3.

The NLP algorithm was constructed in Python 3.7 using the spaCy toolkit [45]. The MetaMap-14 [15,16] engine was installed on our Ubuntu server and accessed from our custom-written Python code using its Java API. We have made our NLP pipeline and the annotation tool publicly available on GitHub [46].

As shown in Fig. 3, clinical notes were read by our custom-written Python scripts [47] for pre-processing. Pre-processing was performed using the Python spaCy package to remove white spaces, special characters, and to convert all characters to lowercase. We also used a custom-built lookup table to map pain-related medical acronyms (including "cp": chest pain,"lbp": lower back pain, and "akp" : anterior knee pain). Our pipeline did not handle spelling errors. However, in our training and validation we did not see any mislabeling due to spelling errors. After pre-processing, larger documents were divided into discrete pages with a maximum character limit of 8,000 to fit the character limit of MetaMap's batch processing software. Truncated notes were passed page-by-page to MetaMap via MetaMap's Java API. MetaMap compiled each file as a 'freetext' and segmented it into 'sentences'. Then, each sentence was processed phrase-by-phrase and was mapped to all possible UMLS concepts. Metamap also, assigned a confidence score for each concept indicating how much each UMLS concept was relevant to the phrase [46]. The NegEx [20] algorithm inside MetaMap was used for negation detection to determine whether mentions of pain terms in the corpus were negated.

Each phrase, together with its assigned clinical concepts, their negation statuses, confidence scores, and ICD codes were stored in a temporary text file. Then, these temporary files were read and the clinical concepts from all phrases of a note were concatenated into a single text file. A sample annotated text is presented in Table 14 in the Supplementary Information. Finally, the program read the processed temporary files phrase by phrase and identified all medical concepts with the 'signs and symptoms' UMLS tag. If multiple medical concepts mapped to a phrase, the program selected the concept with the highest confidence score. The program also extracted medical concepts with a UMLS 'pharmacologic substance' semantic tag in order to identify drug-related phrases. These tags were used to remove drug-related phrases such as "take Tylenol for your back pain". All identified clinical concepts were organized into a data table together with ICD concept IDs, UMLS confidence scores, and negation indices. These data tables were passed to the pain classifier for pain analysis.

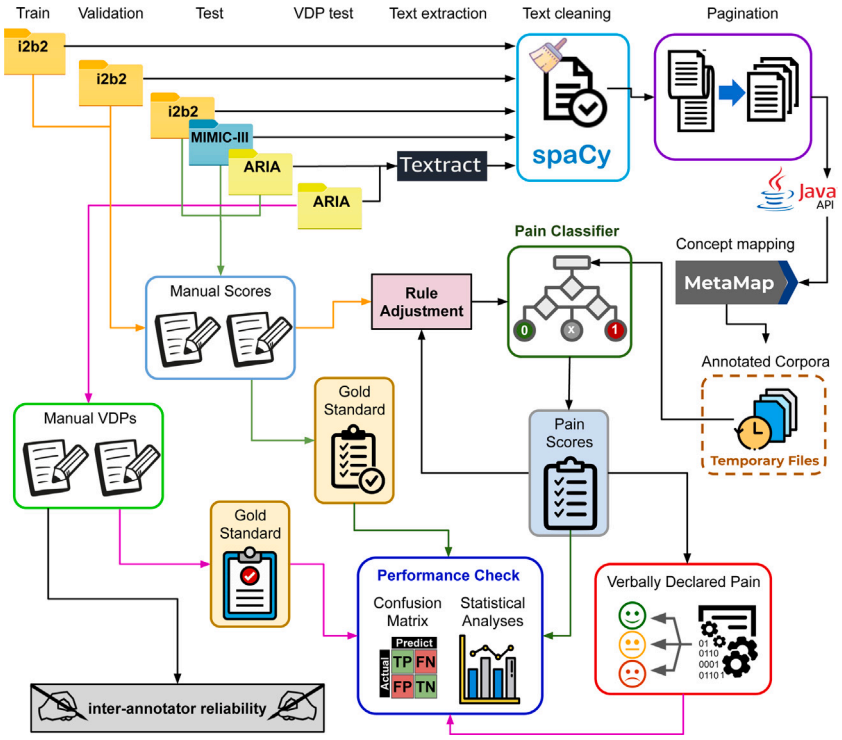

Fig. 3. Our pipeline for medical concept extraction using MetaMap and NegEx. Text from each clinical note was exported as a text document. The Python spaCy package was used for the NLP of patients' consultation notes for text cleaning. The cleaned medical notes were divided into discrete pages (pagination) and passed to the MetaMap and NegEx algorithms via a Java API [16] for the medical name entity tagging and negation detection, respectively. Then, the processed corpora were passed to the pain classifier (Fig. 4) to extract the pain scores. Selection rules were adjusted by evaluating extracted pain scores against the manually annotated pain scores. Finally, the extracted pain scores were stored in the database for VDP calculation, statistical analyses, and performance evaluation.

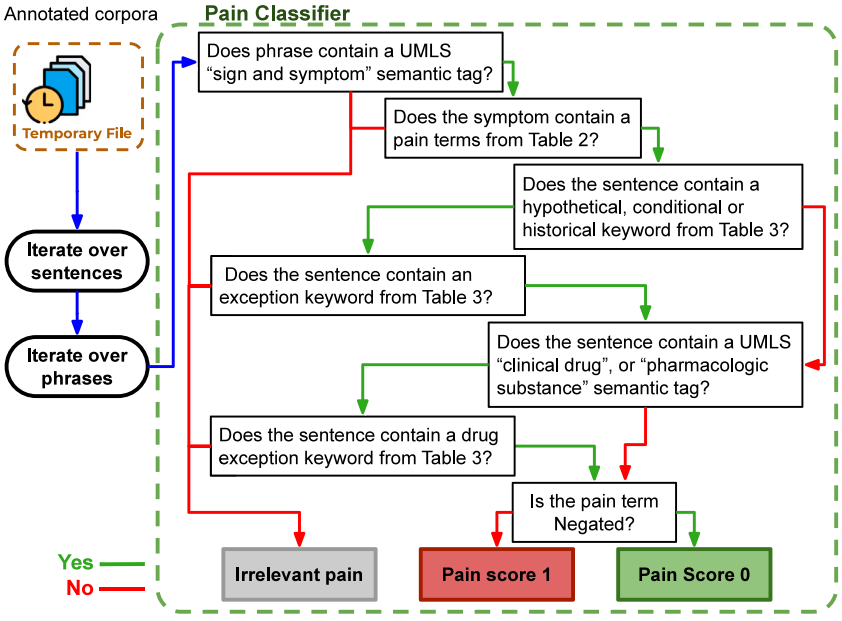

Fig. 4. Our NLP pain classification pipeline to extract the physician-reported pain scores from patients' clinical notes. Annotated files were processed phrase by phrase to filter UMLS 'signs and symptoms' tags and identify pain-related biomedical concepts according to Table 2. Then, sets of rules were developed to remove hypothetical, historical and drug related mentions of the pain and keep the pain term associated to the state of the pain at the time of the hospital visit. Finally, a pain score was assigned to the detected pain term based on the negation status of the phrase.

2.3.2. Step 2: Pain classification

Our rules-based classifier for detecting pain scores is presented in Fig. 4. A lookup table containing Heintzelman et al.'s [33] 66 pain-related medical terminologies was used to determine which 'signs and symptoms', detected by the program, were pain-related (Table 2).

In order to obtain the pain score at the time of the consultation/hospitalization, we excluded irrelevant mentions of pain. For example, we excluded mentions of pain when the patient talked about the history of pain that was not actually presented at the time of the consultation/hospital visit. For this purpose, we trained our pipeline in

Table 2
Pain-related medical terminologies taken from Heintzelman et. al. [33]. These definitions were used to determine pain-related 'signs and symptoms' in the clinical notes by our pain classifier.

Ache	Coccygalgia	Glossalgias	Myodynia	Pressure
Aching	Coccygodynia	Glossodynia	Myosalgia	Proctalgia
Angina	Coccyodynia	Glossodynias	Neuralgia	Rectalgia
Arthralgia	Coccyxdynia	Gonalgia	Neuralgias	Retrosternal
Arthrodynia	Coxalgia	Inguinodynia	Odynophagia	Scapulalgia
Burning	Cp	Lbp	Orchialgia	Scapulodynia
Cephalalgia	Cramp	Low back syndrome	Orchidalgia	Sciatica
Cephalgia	Discomfort	Lumbago	Orchidodynia	Sore
Cephalodynia	Dolor	Lumbalgia	Osteodynia	Tender
Cervicalgia	Dorsalgia	Meralgia	Otalgia	Tightness
Cervicodynia	Dorsodynia	Metatarsalgia	Pain	
Claudication	Dysuria	Muscle weakness	Pancreatalgia	
Coccyalgia	Esophagodynia	Myalgia	Postherpetic	
Coccydynia	Glossalgia	Myalgias	Neuralgia	

Table 3
The lookup tables were formed by examining the training corpus and using the GloVe semantic embedding system. These tables were used to exclude phrases with conditional, hypothetical, historical, and drug-related mentions of pain, and to keep sentences with mentions of the patient's current state of pain in our analysis.

Conditionals	If, whether, when, in case of, in case, as needed, return,
Hypothetical	Might, would, could, should, seek, as needed, call, return, possibly, possible, please, because of, p.r.n.
Historical	History, historical, in the past, previous, before, previously, in the last, prior, recent years
Exceptions	Since, present, current, now, where, because of, prevent, manage, diagnosis, control, found, lasted, treated, resolved, comfort, diagnosis, severe, worsening, aggravated, diffuse, severity, increased, score, high, mild, moderate
Drug mentioned	Clinical drug, pharmacologic substance
Drug exceptions	f-, his, lead, histidine, prevent, wake, level, helium, dob

four iterations by manually auditing 5,138 randomly-selected sentences from the training corpora:

(1) By randomly examining the training corpora, we created a lookup table containing regular expressions related to conditional, hypothetical, and historical terms. These regular expressions were used to search and exclude any pain term used in a conditional, hypothetical, or historical context (Table 3). We used the first validation set to evaluate the performance of the NLP pipeline in correctly detecting valid pain terms.

(2) By examining the training corpora, we created a lookup table containing regular expressions describing current events or ongoing situations such as 'present', 'where', and 'control'. This table (called exceptions) is used to avoid the removal of pain terms related to the current state of the illness. Improvements in the performance of our NLP pipeline was evaluated using the validation set 2.

(3) We used the Global Vectors Word Representation (GloVe) algorithm [48] to generate semantic embedding vectors for all keywords (regular expressions) in the above-mentioned lookup tables. Then, for each keyword, we found five nearest GloVe words in semantic space and added them into the corresponding lookup tables (Table 3). The validation set 3 was used to check the performance of the NLP pipeline at this iteration.

(4) We removed pain terms associated with pain medications by excluding phrases containing the UMLS 'pharmacologic substance' and 'clinical drug' semantic type. Also, by randomly examining the training corpus, we created an exception lookup table to avoid removing pain terms associated with non-pain-related or ambiguous pharmacologic substances like 'dob', 'his', and 'lead'. We used the validation set 4 to evaluate how this iteration improved the performance of the NLP pipeline.

In each iteration of the training, depending on the performance of our NLP pipeline on the validation corpora, we either added more keywords to each of the four lookup tables (Table 3) or removed some keywords from the tables. For example, the keyword 'since' was initially in the conditionals lookup table. But after iteration 1, we moved this keyword to the exceptions lookup table, because, we found that most of the sentences with the keyword 'since' were indicating an ongoing event. We added the keyword 'p.r.n' to the conditionals lookup table, since we found that sentences that includes the 'p.r.n' keyword were most likely talking about a prescription drug. Another example was ambiguous drug names. For instance, we found that the MetaMap classified keyword 'his' as Histidine [Pharmacologic Substance] and keyword 'dob' as Dimethoxybromoamphetamine [Pharmacologic Substance]. We added both these terms to the exception look up table.

Table 4
The confusion matrix for the three-label pain classifiers contains 3 correctly-predicted labels and 6 incorrectly-predicted labels. T_{PP}, T_{NN}, and T_{II} are numbers of sentences that are correctly predicted as score 1, score 0, and irreverent pains, respectively. F_{NP}, F_{PN}, F_{PI}, F_{NI}, F_{IP}, and F_{IN} are numbers of mislabeled sentences.

True label	Predicted label		
	Score 1	Score 0	Irrelevant
Score 1	T_{PP}	F_{PN}	T_{PI}
Score 0	F_{NP}	T_{NN}	F_{NI}
Irrelevant	F_{IP}	F_{IN}	T_{II}

Once satisfied with the training and validation, we did no more development on our NLP pipeline and used gold-standard corpora to evaluate the final performance of the NLP pipeline. Table 3 contains the final versions of the lookup tables. Our NLP pipeline is available as open-source in Ref. [46]

As illustrated in Fig. 4, after passing through the selection rules each phrase was assigned to one of the three scores: valid mention of experienced pain (pain score = 1), valid explicit denial of pain (pain score = 0), and no/irrelevant mention of pain (score = nan) by our pipeline. The third label was primarily used for NLP performance evaluation. Examples of NLP extracted pain scores from i2b2 corpora are provided in Tables 5 and 6.

Table 5
Examples of the sentences from i2b2 corpora that were labeled correctly.

Sentence	Manual pain score	NLP pain score
He states the feeling returned and then persisted, took a 2nd nitro but it only decreased the pain to a [**2192-2-16**].	1	1
Per notes, her abdominal exam was significant for epigastric and right upper quadrant tenderness;	1	1
The patient took one sublingual nitro at home with some relief , but the pain came back as she walking around her home looking for her hospital identification care.	1	1
He had no chest pain but did have diaphoresis and mild nausea and vomiting as well as lightheadedness and some palpitations lasting approximately one hour in duration. ia.	0	0
He had no further episodes of chest pain while in the hospital.	0	0
Patient denies shortness of breath , chest pressure , or syncope.	0	0
He denies fevers or chills, shortness of breath or abdominal pain.	0	0
In July , 1989 , he developed chest pain and suffered an inferior myocardial infarction.	–	–
One week prior to admission , the patient had chest pain , which was quickly relieved by one sublingual nitroglycerin.	–	–
Morphine 15 mg tablet sustained release sig: one (1) tablet sustained release po every 4-6 h as needed for pain.	–	–
If you develop chest pain, nausea, vomiting, throat tightness, clamminess or shortness of breath, call your pcp or go to the emergency room.	–	–

Table 6
Examples of the mislabeled sentences from i2b2 corpora.

Sentence	Manual pain score	NLP pain score
She refused any consultation at this time by the [*** ****] hospital pain service.	-	1
His left groin was not accessed given his c/o left leg pain post surgery 2 months ago.	-	0
The surgical sites were without any exudate or signs of infection and his tenderness in his right upper extremity was markedly decreased.	1	0
In the ambulance , the patient continued to have the pain and she received one more sublingual nitroglycerin and nasal cannula oxygen.	1	-
The patients abdominal pain could be related to intestinal angina.	1	-
asa , o2 , bb , 1 inch of nitropaste for elev bpof note , pt c/o pain on the r mid-lower back which has been present x 1 wk , reproducible on light palpation.	1	-
History of present illness: 74 y/o female with pmh significant for copd, cad, and hypertension admitted to [**hospital1 18**] on [**6-14**] to the surgery service with two days of epigastric and right upper quadrant pain.	1	-
She does however complain of some urinary frequency (on lasix) in the last few days with out any dysuria or urgency.	0	-

2.3.3. Step 3: VDP classification method

Valid pain scores were averaged for each note using Eq. (1) to obtain the average pain intensity at the time of the consultation.

$$Average\ Pain\ Intensity = \frac{\sum(score\ 1\ pains) - \sum(score\ 0\ pains)}{\sum(score\ 1\ pains) + \sum(score\ 0\ pains)} \quad (1)$$

To the best of our knowledge, there are no clinical guidelines to assign a VDP score for overall pain [49]. Our rationale for using a weighted average was to take into account the effect of the number of pain mentions. Also, using a weighted average made it easier for us to map average intensity to VDP. Such a weighted averaging has been previously proposed in the literature for the evaluation of multi-site pain [49–51].

A weighted average pain intensity can range from −1 (when 100% of the valid pain mentions were negated) to 1 (if none of the valid pain mentions was negated). We grouped the average pain intensities in two VDPs by setting the intensity threshold at zero as 'no pain' ($average\ pain\ intensity \leq 0$), and 'pain' ($average\ pain\ intensity \geq 0$). We used the receiver operating characteristic curve (ROC curve) and the area under the curve (AUC) value [52] to examine the performance of a VDP assignment at various intensity thresholds.

2.4. Assessment of the pipeline's performance

Annotated notes from the validation corpora were used to check and tune our NLP pipeline at each iteration of the training. The gold-standard corpora, explained earlier, were used to check the performance of our fully developed NLP pipeline. Confusion matrices were produced to compare the pipeline's performance against expert-annotated gold-standard corpora. To avoid bias, NLP developers were kept blinded to the test corpora throughout the entire process.

The confusion matrix for our three-label pain classifier is a 3 × 3 matrix, as presented in Table 4. This matrix includes 3 TRUE labels for correctly-scored sentences, and 6 FALSE labels for incorrectly-scored sentences (more details are provided in section 7.5).

We evaluated the performance of our NLP pipeline for pain scoring and VDP assignment by calculating the precision, recall, and F1-score (F1) from the confusion matrices [52].

3. Results

3.1. Pain classifier

We tested our NLP pipeline's ability to extract pain terms from notes in the i2b2, MIMIC-III, and ARIA corpora. By processing all the available corpora, we found 19,851, 12,071, and 1883 suggested pain concepts, respectively. Note that these pain concepts include all the

Table 7
The frequency of score 0 and score 1 pain terms labeled by the NLP pipeline in each of the three corpora. Total number of valid pain terms are provided inside the brackets.

	i2b2 % (n = 4385)	MIMIC-III % (n = 3109)	ARIA % (n = 2572)
Score 1 pain	64.9	54.1	78.3
Score 0 pain	35.1	45.9	21.7

Table 8
Verbally-declared pain (VDP) at the time of the consultation using all available notes from each corpora. The VDP was obtained by averaging over all pain scores in each note. Percentile values are specified in parentheses.

	i2b2	MIMIC-III	ARIA
Pain	706 (56.5%)	442 (50.4%)	511 (64.9%)
No pain	305 (24.4%)	262 (29.9%)	104 (13.2%)
No mention of pain	238 (19.1%)	173 (19.7%)	173 (21.9%)

UMLS concepts that were extracted from the clinical notes. This means that multiple pain concepts may have been mapped to one phrase as described earlier.

The result of our rule-based pain detection pipeline (shown in Fig. 4) for detecting the pain score is presented in Table 7. Using the UMLS confidence score to remove duplicate concepts, we obtained uniquely-mapped experienced pain terms and explicitly denied pain terms from the i2b2, MIMIC-III, and ARIA corpora. Finally, by removing conditional, hypothetical, and drug-related pain terms, we obtained 2845, 1682, and 2013 relevant terms presenting the pain score 1 as well as 1540, 1427, and 559 score 0 pain terms in the i2b2, MIMIC-III, and ARIA corpora, respectively. Table 5 contains a few example sentences from i2b2 corpora in which pain scores were correctly labeled by our NLP pipeline. Examples of pain terms that were not labeled correctly by our NLP pipeline are provided in Table 6.

On averaging over the pain scores in each note using Eq. (1), we obtained the VDP at the time of consultation/hospitalization in the three corpora. Distribution of the VDP is presented in Table 8. Based on our VDP calculations, we found that pain was not documented in 22% of the cancer notes, 13% of our cancer patients denied the experience of pain and at least 65% of cancer patients experienced some level of pain. These results were in agreement with the results reported in the other papers [53].

3.2. Inter-annotator agreement

Inter-annotator agreement among 6 annotators in assigning notes to a 5-grade pain scale is provided in Table 18 (Supplementary Information). We calculated Fleiss' kappa measure and obtained a moderate agreement among 6 annotators ($\kappa = 0.43$). This indicated that pain scores were not sufficiently documented in the consultation notes. Therefore, we instead defined a 3-grade pain scale (called VDP status) by merging 'mild', 'moderate' and 'severe pain' assignments into a single category as 'pain'. We measured the inter-annotator agreement again and we obtained substantial agreement between six annotators in assigning VDP with Fleiss' kappa measure of $\kappa = 0.66$.

3.3. Performance of the pain classifier

The confusion matrices, generated by comparing NLP-extracted pain scores against expert-annotated gold-standard from each corpus, are presented in Table 9. Based on these confusion matrices, we calculated precision, recall and F1-score. These results are summarized in Table 10. To compare the performance of our NLP pipeline with the prior studies, we provided the performances of the pain extraction NLP algorithms presented by Heintzelman et al. [33] and Bui and Zeng-Treitler [35].

Table 9
Following the approach described in Table 4, for each corpus a three-class confusion matrix was obtained. The name of the corresponding corpus is mentioned in the top left cell of the matrix.

True label	Predicted label		
i2b2	Pain score 1	Pain score 0	Irrelevant
Pain score 1	78	1	11
Pain score 0	0	22	1
Irrelevant	5	1	2241
True label	Predicted label		
MIMIC-III	Pain score 1	Pain score 0	Irrelevant
Pain score 1	51	1	6
Pain score 0	0	15	3
Irrelevant	3	1	2635
True label	Predicted label		
ARIA	Pain score 1	Pain score 0	Irrelevant
Pain score 1	70	1	13
Pain score 0	1	24	5
Irrelevant	10	1	1007

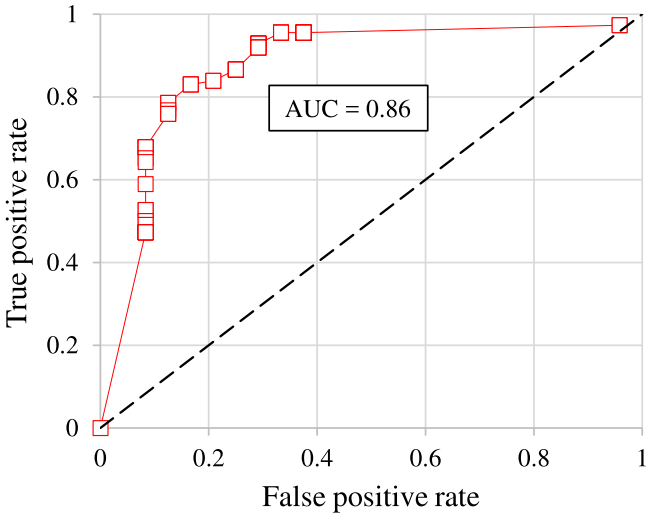

Fig. 5. The ROC curve was generated to investigate the performance of a VDP classification method at various intensity thresholds. The AUC is calculated to be 0.86.

3.4. Performance of the VDP classifier

The performance of the VDP classification method was evaluated using the 3-grade VDP gold-standard corpora. A 3 × 3 confusion matrix was formed for the three-grade VDP, as explained in Section 2.4. Table 11 shows the confusion matrix for NLP extracted VDP. The ROC curve is plotted in Fig. 5 for various intensity thresholds. The AUC is calculated to be 0.86.

Of the 150 notes selected for the performance evaluation, 14 notes did not have any valid mention of pain (no mention of pain), 112 notes had 'pain', and 24 had 'no pain' (denied pain) VDP. Among the 112 notes with the mentions of experienced pain, our VDP extraction method correctly classified 104 of them while five were misclassified as no pain and the other three were misclassified as no mention of pain. Among the 24 notes with no pain VDP, our pipeline correctly classified 16 of them and incorrectly labeled seven as pain and one as no mention of pain.

Based on these results, we calculated the precision, recall, and F1-score for the VDP extraction method that are shown in Table 12. We achieved 92%, 76%, and 75% precision in classifying the notes into the 'pain', 'no pain', and 'no mention of pain' VDP, respectively.

Table 10
The precision (P), recall (R) and F1-score (F) of the pain detection pipeline calculated based on the confusion matrices presented in Table 9. The performances of the NLP pipelines from prior studies are provided for a comparison.

Author	Pain score 1			Pain score 0		
	P	R	F	P	R	F
Present study (i2b2)	94.0	86.7	90.2	91.7	95.7	93.6
Present study (MIMIC-III)	94.4	88.0	91.1	88.2	83.3	85.7
Present study (ARIA)	86.4	83.3	84.8	92.3	80.0	85.7
Heintzelman et al. [33][a]	86	95	90	82	95	88
Bui and Zeng-Treitler [35][b]	73.2	56.6	63.8	78.8	74.2	76.4

[a]Calculated based on the manual annotation of 111 pain mentions that were extracted from 30 discharge summaries from i2b2 database.
[b]Calculated based on manual annotation of 702 pain mentions that were extracted from 100 documents from the US Department of Veterans Affairs' (VA) electronic medical records.

Table 11
Following the approach described in Section 2.4, a three-point VDP confusion matrix was formed based on the manual audit of 120 randomly-selected notes from the ARIA corpora.

True VDP	Predicted VDP		
ARIA	Pain	No pain	No mention of pain
Pain	104	5	3
No pain	7	16	1
No mention of pain	2	0	12

Table 12
The precision, recall and F1-score of the VDP extraction method has been calculated using Table 11.

ARIA	Precision	Recall	F1-score
Pain	92.0%	92.9%	92.4%
No pain	76.2%	66.7%	71.4%
No mention of pain	75.0%	85.7%	80.4%

4. Discussion

4.1. Quality of corpora

Comparing the number of words and sentences in Fig. 1, we found that the consultation notes from the ARIA corpus contained noticeably fewer words and sentences compared to the discharge summaries from the i2b2 and MIMIC-III corpora. Since notes from the i2b2 and MIMIC-III corpora were pre-processed and de-identified for public use, they contained more broken sentences. Nonetheless, we found that the distribution of the length of words and sentences were similar across all three corpora. Therefore, the similarity of the datasets was not very affected by the pre-processing and de-identification steps. This suggests that notes from various resources are similar enough to be used together in a study such as this.

4.2. Distribution of the pain terms in the notes

Distribution of the pain terms in the notes from three corpora, presented in Table 19 in the Supplementary Information, revealed that pain distribution from the ARIA corpus was notably different from the other two corpora. As expected, the ARIA corpus included only patients with bone metastases, hence, there were more mentions of bone-related pain terms such as back pain and pelvic pain. We also observed that almost 58% of the experienced pain was reported as generic pain without specifying the pain site in the ARIA corpus while this was only 34% and 38% in the other two corpora. We suspect that it was because the consultation notes in ARIA were prepared by radiation oncologists who solely examined cancer patients, while discharge summaries were prepared by general physicians who visited patients with a variety of conditions.

Comparing the experienced pain with the total pain mentions, we detected that pain was experienced in 65% and 54% of the cases in the i2b2 and MIMIC-III corpora respectively, while this number increased to 78% in the ARIA corpus. Again, we assume that the explanation for this might be due to the nature of these three corpora with i2b2 and MIMIC-III containing notes for patients visiting general hospitals while our ARIA database included exclusively notes for cancer patients with bone metastases. Remarkably, the 78% experienced pain for metastatic cancer patients agrees with the results reported in several other studies [54,55].

4.3. Accuracy of the pain score measurements

Performance of our NLP pipeline was evaluated using the gold-standard test sets explained in Section 2.4. As presented in Table 10, our pipeline outperformed prior pain detection pipelines developed by Heintzelman et al.[33] and Bui and Zeng-Treitler [35].

Once we fully trained and tested our pipeline using the i2b2 training corpus, we examined the generalizability of our NLP pain detection pipeline using independent corpora from MIMIC-III and ARIA. Note that our NLP pipeline was used on the MIMIC-III and ARIA corpora without further training on these corpora. The precision of our NLP pipeline in detecting score 1 pain did not change when we applied it to the MIMIC-III corpora. However, it dropped to 86% when we applied our pipeline to the ARIA corpora. The reason for having more mislabeled score 1 pain in the ARIA corpus can be attributed to the difference in the corpus type. The i2b2 and MIMIC-III corpora were general hospital discharge summary notes, while the ARIA corpus comprises radiation oncology consultation notes. As stated previously, up to 50% reduction in the precision is commonly expected when moving from public corpora to private corpora. Therefore, a 12% drop in the precision of our NLP pipeline was reasonable. This suggests that NLP pipelines that are trained on one type of documents (i.e. hospital discharge summaries in this case) can be successfully transferred to analyze patients' other clinical notes (such as cancer consultation notes in this study).

The precision in detecting score 0 pain reduced from 92% to 88% when the MIMIC-III corpus was analyzed. The decrease in precision might be as a result of more diverse negation terms in the MIMIC-III, which includes notes from more diverse sources compared to the i2b2 database. The precision of our pipeline in detecting score 0 pain terms was 92% when analyzing the ARIA corpus. The main reason for such a high precision was because of better sentence segmentation in ARIA corpus compared to i2b2 and MIMIC-III corpora. Both the i2b2 and MIMIC-III were de-identified corpora with a lot of broken sentences. Therefore, it was much harder for our NLP pipeline to detect negation (score 0 pain terms). Examples of mislabeled pain terms are presented in Table 6 in the Supplementary Information.

The recall parameter provided more information about the behavior of our NLP system. Recall was the measure of how well our pipeline correctly identified all true labels. In the i2b2 and MIMIC-III corpora, we achieved 87% and 88% recall in detecting score 1 pain, respectively. The recall decreased to 83% for the ARIA corpus. As shown in Table 9, in the ARIA corpus, a notable number of the score 1 pain was assigned

as irrelevant pain. We believe this noticeable mislabeling were related to the pain terms that were describing patient's previous experience of having pain. As expected, most of cancer patients had a history of long term chronic pain which presumably made it difficult for our pipeline to separate them from pain at the time of the consultation.

The recall values for detecting all mentions of score 0 pain were 96%, 83%, and 80% for the i2b2, MIMIC-III, and ARIA corpora, respectively. We believe that this variation in the recall values of score 0 pain was partially due to the layout of the notes in each corpus. For example, in the i2b2 corpus that was used to train our NLP pipeline, each note had a separate section for prescription drugs. Therefore, the drug-related pain terms could be filtered much easier than in the MIMIC-III corpus in which the prescription drugs were mentioned within the notes in an unstructured format. It should be noted that , as we explained in section 7.1, we did not cut any segment of the notes in any of the corpora to assure the generalizability of our pipeline.

Having fewer score 0 pain terms might also influence the calculated recall values. Table 9 shows that there were only 23, 18 and 30 score 0 pain terms in i2b2, MIMIC-III and ARIA validation corpora, respectively. This means that any mislabeled score 0 pain, introduced a large uncertainty to the recall values.

The overall performance of our NLP pipeline on various corpora was also evaluated using F1-scores. The F1-score did not vary much among the three corpora. F1-score of score 1 pain only decreased from 0.90 in i2b2 to 0.85 in ARIA corpus. Similarly, F1-score of score 0 pain changed from 0.94 in i2b2 to 0.86 in ARIA corpus.

4.4. Accuracy of the VDP extraction

Based on the ROC curve with an AUC value of 0.86 (Fig. 5), our VDP extraction method had good performance in distinguishing between patients with and without pain. As presented in Table 12, our VDP extraction method successfully detected 'pain' with 92.0% precision and 92.9% recall. However, it showed fundamental limitations in detecting 'no pain', with 76.2% precision and 66.7% recall. The main reason for such a high recall and low precision in detecting no pain VDP was that ARIA was an imbalanced corpus, where the classes were not represented equally (i.e. there was ~ 5x more experienced pain than no pain cases, as shown in Table 11.)

In addition, investigating the notes in the i2b2 training set, we noticed that when patients reported pain at multiple sites in their body, our classification method was not able to extract VDP precisely. Our method of measuring VDP was confounded by the reality of the notes of metastatic cancer patients, because, for these patients, it is expected to have multiple pain sites with different pain scores in each site.

One possible solution is to add functionality to obtain pain severity from patients' consultation notes by analyzing the pain assessment terminologies (such as severe, mild, controlled) and by capturing numerical pain scores for each identified pain site directly from the consultation notes.

5. Conclusion

Our database-independent NLP pipeline, trained using i2b2 hospital discharge summary corpora, was successfully implemented to detect and classify pain from the publicly-available MIMIC-III hospital discharge summary corpus, and our institutional radiation oncology ARIA consultation note database for cancer patients with bone metastases. The pipeline's performance was evaluated against physician-annotated gold standard corpora. Our pipeline achieved a precision and a recall of 89% and 82% in detecting physician-reported pain, respectively, demonstrating successful and state-of-the-art extraction and classification of pain from radiation oncology clinical notes. It also automatically assigned a VDP for each clinical note with 84% and 80% overall precision and recall.

An important and intended application of our NLP tool is that it can be used to reliably extract physician-reported cancer pain from clinical notes in radiation oncology, where the pain is not otherwise documented through structured data entry. Having access to this database-independent NLP pain-extraction pipeline will facilitate further informatics and data-mining studies in radiation oncology that require access to pain information that is typically very difficult to obtain.

CRediT authorship contribution statement

Hossein Naseri: Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Visualization, Writing – original draft. **Kamran Kafi:** Validation, Writing – review & editing. **Sonia Skamene:** Validation. **Marwan Tolba:** Validation. **Mame Daro Faye:** Validation. **Paul Ramia:** Validation. **Julia Khriguian:** Validation. **John Kildea:** Conceptualization, Investigation, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Funding for this research was provided by the startup grant of Dr. John Kildea at Research Institute of the McGill University Health Centre (RI-MUHC), RI-MUHC studentship, Ruth and Alex Dworkin scholarship award from the McGill University - Faculty of Medicine, and Grad Excellence Award-00293 from the McGill University - Department of Physics. The authors would like to thank Ms. Haley Patrick for the manual audit of our validation sets. We thank Mr. Farzin Khosrow-Khavar for his help with sample size evaluation and statistical analysis. We also thank Dr. Marc David for his clinical support.

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jbi.2021.103864.

References

[1] World Health Organization, WHO Guidelines for the Pharmacological and Radiotherapeutic Management of Cancer Pain in Adults and Adolescents, World Health Organization, 2018, http://www.ncbi.nlm.nih.gov/pubmed/30776210.

[2] M.G. Nayak, A. George, M.S. Vidyasagar, S. Mathew, S. Nayak, B.S. Nayak, Y.N. Shashidhara, A. Kamath, Quality of life among cancer patients, Ind. J. Palliat. Care 23 (4) (2017) 445–450, http://dx.doi.org/10.4103/IJPC.IJPC_82_17.

[3] L.S. Simon, Relieving pain in america: a blueprint for transforming prevention, care, education, and research, J. Pain Palliat. Care Pharmacother. 26 (2) (2012) 197–198, http://dx.doi.org/10.3109/15360288.2012.678473.

[4] G.G. Page, S. Ben-Eliyahu, The immune-suppressive nature of pain, Sem. Oncol. Nursing 13 (1) (1997) 10–15, http://dx.doi.org/10.1016/S0749-2081(97)80044-7.

[5] D.B. Gordon, J.L. Dahl, C. Miaskowski, B. McCarberg, K.H. Todd, J.A. Paice, A.G. Lipman, M. Bookbinder, S.H. Sanders, D.C. Turk, D.B. Carr, American Pain Society Recommendations for improving the quality of acute and cancer pain management: American pain society quality of care task force, Arch. Internal Med. 165 (14) (2005) 1574–1580, http://dx.doi.org/10.1001/archinte.165.14.1574.

[6] M.P. Cadogan, J.F. Schnelle, N.R. Al-Sammarrai, N. Yamamoto-Mitani, G. Cabrera, D. Osterweil, S.F. Simmons, A standardized quality assessment system to evaluate pain detection and management in the nursing home, J. Amer. Med. Direct. Assoc. 6 (1) (2005) 1–9, http://dx.doi.org/10.1016/j.jamda.2004.12.002.

[7] T.J. Keay, The mind-set of pain assessment, J. Amer. Med. Direct. Assoc. 6 (1) (2005) 77–78, http://dx.doi.org/10.1016/j.jamda.2004.12.011.

[8] L. Ohno-Machado, Realizing the full potential of electronic health records: the role of natural language processing, J. Amer. Med. Inform. Assoc. 18 (5) (2011) http://dx.doi.org/10.1136/amiajnl-2011-000501, 539–539.

[9] S. Bird, E. Klein, E. Loper, Natural Language Processing with Python, first ed., OŕEilly Media, Inc., 2009.

[10] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.

[11] WHO, International classification of diseases, 11th revision (ICD-11), WHO (2019) http://www.who.int/classifications/icd/en/.

[12] S. International, SNOMED CT January 2020 International Edition - SNOMED International Release notes - SNOMED International Release Management - SNOMED Confluence, https://confluence.ihtsdotools.org/display/RMT, [online].

[13] O. Bodenreider, The unified medical language system (UMLS): integrating biomedical terminology, Nucleic Acids Res. 32 (Database issue) (2004) 267–270, http://dx.doi.org/10.1093/nar/gkh061.

[14] Bethesda (MD), UMLS® Reference Manual, National Library of Medicine (US), 2009, https://www.ncbi.nlm.nih.gov/books/NBK9676/.

[15] A.R. Aronson, F.M. Lang, An overview of metamap: Historical perspective and recent advances, J. Amer. Med. Inform. Assoc. 17 (3) (2010) 229–236, http://dx.doi.org/10.1136/jamia.2009.002733.

[16] D. Demner-Fushman, W.J. Rogers, A.R. Aronson, Metamap lite: an evaluation of a new java implementation of metamap, J. Amer. Med. Inform. Assoc. 24 (4) (2017) 841–844, http://dx.doi.org/10.1093/jamia/ocw177.

[17] J. Zhang, X. Long, T. Suel, Performance of compressed inverted list caching in search engines, in: Proceeding of the 17th International Conference on World Wide Web 2008, WWW'08, ACM Press, New York, New York, USA, 2008, pp. 387–396, http://dx.doi.org/10.1145/1367497.1367550.

[18] L.M. Simon, S. Karg, A.J. Westermann, M. Engel, A.H.A. Elbehery, B. Hense, M. Heinig, L. Deng, F.J. Theis, Metamap: an atlas of metatranscriptomic reads in human disease-related RNA-seq data, GigaScience 7 (6) (2018) http://dx.doi.org/10.1093/gigascience/giy070.

[19] R. Reátegui, S. Ratté, Comparison of metamap and cTAKES for entity extraction in clinical notes, BMC Med. Inform. Decis. Mak. 18 (2018) http://dx.doi.org/10.1186/s12911-018-0654-2.

[20] W.W. Chapman, W. Bridewell, P. Hanbury, G.F. Cooper, B.G. Buchanan, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (5) (2001) 301–310, http://dx.doi.org/10.1006/jbin.2001.1029.

[21] S. Wu, T. Miller, J. Masanz, M. Coarr, S. Halgrim, D. Carrell, C. Clark, Negations not solved: Generalizability versus optimizability in clinical natural language processing, in: C. Lovis (Ed.), PLoS ONE 9 (11) (2014) e112774, http://dx.doi.org/10.1371/journal.pone.0112774.

[22] Z. Zeng, Y. Deng, X. Li, T. Naumann, Y. Luo, Natural language processing for EHR-based computational phenotyping, IEEE/ACM Trans. Comput. Biol. Bioinform. 16 (1) (2019) 139–153, http://dx.doi.org/10.1109/TCBB.2018.2849968.

[23] X. Wang, A. Chused, N. Elhadad, C. Friedman, M. Markatou, Automated knowledge acquisition from clinical narrative reports, in: AMIA ... Annual Symposium Proceedings / AMIA Symposium. AMIA Symposium, 2008 (2008) 783–787.

[24] I.V. Haller, C.M. Renier, M. Juusola, P. Hitz, W. Steffen, M.J. Asmus, T. Craig, J. Mardekian, E.T. Masters, T.E. Elliott, Enhancing risk assessment in patients receiving chronic opioid analgesic therapy using natural language processing, Pain Med. 18 (10) (2016) 1952–1960, http://dx.doi.org/10.1093/pm/pnw283.

[25] T.A. Koleck, C. Dreisbach, P.E. Bourne, S. Bakken, Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review, J. Amer. Med. Inform. Assoc. 26 (4) (2019) 364–379, http://dx.doi.org/10.1093/jamia/ocy173.

[26] A. Hardjojo, A. Gunachandran, L. Pang, M.R.B. Abdullah, W. Wah, J.W.C. Chong, E.H. Goh, S.H. Teo, G. Lim, M.L. Lee, W. Hsu, V. Lee, M.I.-C. Chen, F. Wong, J.S.K. Phang, Validation of a natural language processing algorithm for detecting infectious disease symptoms in primary care electronic medical records in Singapore, JMIR Med. Inform. 6 (2) (2018) e36, http://dx.doi.org/10.2196/medinform.8204.

[27] G.K. Savova, E. Tseytlin, S. Finan, M. Castine, T. Miller, O. Medvedeva, D. Harris, H. Hochheiser, C. Lin, G. Chavan, R.S. Jacobson, Deepphe: A natural language processing system for extracting cancer phenotypes from clinical records, Cancer Res. 77 (21) (2017) e115–e118, http://dx.doi.org/10.1158/0008-5472.CAN-17-0615.

[28] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearb. Med. Inform. (2008) 128–144, http://dx.doi.org/10.1055/s-0038-1638592.

[29] S.S. Pakhomov, H. Hemingway, S.A. Weston, S.J. Jacobsen, R. Rodeheffer, V.L. Roger, Epidemiology of angina pectoris: Role of natural language processing of the medical record, Amer. Heart J. 153 (4) (2007) 666–673, http://dx.doi.org/10.1016/j.ahj.2006.12.022.

[30] W.K. Tan, S. Hassanpour, P.J. Heagerty, S.D. Rundell, P. Suri, H.T. Huhdanpaa, K. James, D.S. Carrell, C.P. Langlotz, N.L. Organ, E.N. Meier, K.J. Sherman, D.F. Kallmes, P.H. Luetmer, B. Griffith, D.R. Nerenz, J.G. Jarvik, Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain, Acad. Radiol. 25 (11) (2018) 1422–1432, http://dx.doi.org/10.1016/j.acra.2018.03.008.

[31] T.Y. Tian, I. Zlateva, D.R. Anderson, Using electronic health records data to identify patients with chronic pain in a primary care setting, J. Amer. Med. Inform. Assoc. 20 (E2) (2013) e275, http://dx.doi.org/10.1136/amiajnl-2013-001856.

[32] S.J. Fodeh, D. Finch, L. Bouayad, S.L. Luther, H. Ling, R.D. Kerns, C. Brandt, Classifying clinical notes with pain assessment using machine learning, Med. Biol. Eng. Comput. 56 (7) (2018) 1285–1292, http://dx.doi.org/10.1007/s11517-017-1772-1.

[33] N.H. Heintzelman, R.J. Taylor, L. Simonsen, R. Lustig, D. Anderko, J.A. Haythornthwaite, L.C. Childs, G.S. Bova, Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text, J. Amer. Med. Inform. Assoc. 20 (5) (2013) 898–905, http://dx.doi.org/10.1136/amiajnl-2012-001076.

[34] A.S. Eisman, N.R. Shah, C. Eickhoff, G. Zerveas, E.S. Chen, W.-C. Wu, I.N. Sarkar, Extracting angina symptoms from clinical notes using pre-trained transformer architectures, 2020, arXiv:2010.05757.

[35] D.D.A. Bui, Q. Zeng-Treitler, Learning regular expressions for clinical text classification, J. Amer. Med. Inform. Assoc. 21 (5) (2014) 850–857, http://dx.doi.org/10.1136/amiajnl-2013-002411.

[36] V. Major, A. Surkis, Y. Aphinyanaphongs, Utility of general and specific word embeddings for classifying translational stages of research, in: AMIA. Annual Symposium Proceedings. AMIA Symposium, vol. 2018, NLM (Medline), 2018, pp. 1405–1414.

[37] Y. Si, J. Wang, H. Xu, K. Roberts, Enhancing clinical concept extraction with contextual embeddings, J. Amer. Med. Inform. Assoc. 26 (11) (2019) 1297–1304, http://dx.doi.org/10.1093/jamia/ocz096.

[38] C. Tao, M. Filannino, O. Uzuner, Prescription extraction using CRFs and word embeddings, J. Biomed. Inform. 72 (2017) 60–66, http://dx.doi.org/10.1016/j.jbi.2017.07.002.

[39] D.T. Heinze, M.L. Morsch, B.C. Potter, R.E. Sheffer, Medical i2b2 NLP smoking challenge: The A-life system architecture and methodology, J. Amer. Med. Inform. Assoc. 15 (1) (2008) 40–43, http://dx.doi.org/10.1197/jamia.M2438.

[40] O. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Amer. Med. Inform. Assoc. 14 (5) (2007) 550–563, http://dx.doi.org/10.1197/jamia.M2444.

[41] A.E. Johnson, T.J. Pollard, L. Shen, L.W.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data 3 (1) (2016) 1–9, http://dx.doi.org/10.1038/sdata.2016.35.

[42] D. Malmgren, Textract documentation release 1.1.0, 2014, https://textract.readthedocs.io/en/stable/.

[43] W.G. Cochran, Sampling Techniques, third ed., John Wiley, 1977.

[44] J.L. Fleiss, J. Cohen, The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability, Educ. Psychol. Meas. 33 (1973) 613–619, http://dx.doi.org/10.1177/001316447303300309.

[45] spaCy, · Industrial-strength Natural Language Processing in Python, https://spacy.io/ [online].

[46] H. Naseri, Textractor; tools for pain scoring, in: GitHub Repository, GitHub, 2021, http://dx.doi.org/10.5281/zenodo.4649625.

[47] Intro to data structures - pandas 1.0.5 documentation [online].

[48] J. Pennington, R. Socher, C.D. Manning, GloVe: Global Vectors for Word Representation, https://nlp.stanford.edu/pubs/glove.pdf [online].

[49] M.S. Wallace, J. North, E.J. Grigsby, L. Kapural, M.R. Sanapati, S.G. Smith, C. Willoughby, P.J. McIntyre, S.P. Cohen, R.M. Rosenthal, S. Ahmed, R. Vallejo, F.M. Ahadian, T.L. Yearwood, A.W. Burton, E.J. Frankoski, J. Shetake, S. Lin, B. Hershey, B. Rogers, N. Mekel-Bobrov, An integrated quantitative index for measuring chronic multisite pain: The multiple areas of pain (MAP) study, Pain Med. (United States) 19 (2018) 1425–1435, http://dx.doi.org/10.1093/pm/pnx325, https://pubmed.ncbi.nlm.nih.gov/29474648/.

[50] S.D. Rundell, K.V. Patel, M.A. Krook, P.J. Heagerty, P. Suri, J.L. Friedly, J.A. Turner, R.A. Deyo, Z. Bauer, D.R. Nerenz, A.L. Avins, S.S. Nedeljkovic, J.G. Jarvik, Multi-site pain is associated with long-term patient-reported outcomes in older adults with persistent back pain, Pain Med. (United States) 20 (2019) 1898–1906, http://dx.doi.org/10.1093/pm/pny270, https://pubmed.ncbi.nlm.nih.gov/30615144/.

[51] M.P. Jensen, C. Tomé-Pires, E. Solé, M. Racine, E. Castarlenas, R. o de la Vega, J. Miró, Assessment of pain intensity in clinical trials: Individual ratings vs composite scores, Pain Med. (United States) 16 (2015) 141–148, http://dx.doi.org/10.1111/pme.12588, https://pubmed.ncbi.nlm.nih.gov/25280226/.

[52] A. Tharwat, Classification assessment methods, Appl. Comput. Inform. 17 (1) (2018) 168–192, http://dx.doi.org/10.1016/j.aci.2018.08.003.

[53] M. Kuchuk, C.L. Addison, M. Clemons, I. Kuchuk, P. Wheatley-Price, Incidence and consequences of bone metastases in lung cancer patients, J. Bone Oncol. 2 (1) (2013) 22–29, http://dx.doi.org/10.1016/j.jbo.2012.12.004.

[54] A. Tsuya, T. Kurata, K. Tamura, M. Fukuoka, Skeletal metastases in non-small cell lung cancer: A retrospective study, Lung Cancer 57 (2007) 229–232, http://dx.doi.org/10.1016/j.lungcan.2007.03.013.

[55] M. Kuchuk, C.L. Addison, M. Clemons, I. Kuchuk, P. Wheatley-Price, Incidence and consequences of bone metastases in lung cancer patients, J. Bone Oncol. 2 (1) (2013) 22–29, http://dx.doi.org/10.1016/j.jbo.2012.12.004.