Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients

Jean Coquet[a], Selen Bozkurt[a,b], Kathleen M. Kan[c], Michelle K. Ferrari[c], Douglas W. Blayney[a,d], James D. Brooks[c,d], Tina Hernandez-Boussard[a,b,e,*]

[a] *Department of Medicine, Stanford University, Stanford, CA, USA*
[b] *Department of Biomedical Data Science, Stanford University, Stanford, USA*
[c] *Department of Urology, Stanford University School of Medicine, Stanford, USA*
[d] *Stanford Cancer Institute, Stanford University School of Medicine, Stanford, USA*
[e] *Department of Surgery, Stanford University School of Medicine, Stanford, USA*

ARTICLE INFO

Keywords:
Electronic health records
Natural language processing
Machine learning
Prostate cancer

ABSTRACT

Objective: Clinical care guidelines recommend that newly diagnosed prostate cancer patients at high risk for metastatic spread receive a bone scan prior to treatment and that low risk patients not receive it. The objective was to develop an automated pipeline to interrogate heterogeneous data to evaluate the use of bone scans using a two different Natural Language Processing (NLP) approaches.

Materials and Methods: Our cohort was divided into risk groups based on Electronic Health Records (EHR). Information on bone scan utilization was identified in both structured data and free text from clinical notes. Our pipeline annotated sentences with a combination of a rule-based method using the ConText algorithm (a generalization of NegEx) and a Convolutional Neural Network (CNN) method using word2vec to produce word embeddings.

Results: A total of 5500 patients and 369,764 notes were included in the study. A total of 39% of patients were high-risk and 73% of these received a bone scan; of the 18% low risk patients, 10% received one. The accuracy of CNN model outperformed the rule-based model one (F-measure = 0.918 and 0.897 respectively). We demonstrate a combination of both models could maximize precision or recall, based on the study question.

Conclusion: Using structured data, we accurately classified patients' cancer risk group, identified bone scan documentation with two NLP methods, and evaluated guideline adherence. Our pipeline can be used to provide concrete feedback to clinicians and guide treatment decisions.

1. Introduction

Prostate cancer is the most common cancer diagnosed in North American and European men. [1,2] Most prostate cancers are diagnosed by screening practices with low-grade and low stage disease; however, approximately 15% of newly diagnosed cancers carry a high risk of spread and eventual mortality. [3] High risk prostate cancers are defined by a variety of clinical parameters, including clinical stage, prostate-specific antigen (PSA) values, and biopsy Gleason score. [4] Because definitive treatments (surgery or radiation therapy) are accompanied by substantial morbidity risk (including long lasting urinary incontinence and impotence) patients who have a low likelihood of cure (especially those with bone metastases) should not receive potentially morbid procedures. Patients at high risk for bone metastases at their presentation should receive a radionuclide bone scan (hereafter, bone scan) to better inform treatment decisions. [5] However, there is concern regarding the over-use and under-use of bone scans across different risk groups.

Clinical guidelines are used to guide patient and physician decision-making and to ensure patients are offered appropriate, evidence-based, care. Many guidelines include guidance for medical imaging. [6] The National Comprehensive Cancer Network (NCCN) and American Urological Association (AUA) guidelines recommend that patients with advanced stage and local/regional high-risk prostate cancer *receive* a bone scan for staging purposes and that low-risk patients *not receive* a bone scan prior to treatment. [7,8] Despite these largely agreed-upon guidelines, bone scans are often over-used in low-risk patients; a recent study reported that up to 35% of low-risk patients received an

[*] Corresponding author at: 1265 Welch Road, #245, Stanford, CA 94305-5246, USA.
E-mail address: boussard@stanford.edu (T. Hernandez-Boussard).

https://doi.org/10.1016/j.jbi.2019.103184
Received 27 November 2018; Received in revised form 15 April 2019; Accepted 19 April 2019
Available online 20 April 2019

unnecessary bone scan.[9,10] On the other hand, bone scans may be underutilized in high-risk patients, which may subject advanced disease patients to unnecessary morbid and ineffective procedures.[5,11] Developing methods to systematically evaluate guideline adherence is essential to assess and improve health care quality.

Clinical features needed to appropriately classify patients into low and high risk categories are embedded in multiple data sources and scattered throughout electronic health records (EHRs) and manual review of information contained within free-text formats is time-consuming and expensive. [12,13] Given the complexity of assigning prostate cancer patients into 'low risk' and 'high risk' categories, automated methods are needed to extract and synthesize the clinical data. Natural Language Processing (NLP) methods represent a solution that can aid in extracting information from provider notes to answer pertinent clinical questions for health outcomes research. [14] Different approaches exist, some based on lexical and linguistic rules [15,16] and others based on machine learning approaches, [17,18] and recently there is a recent strong interest in using deep learning methods for knowledge extraction. [19] In addition, the use of hybrid methods that combine these approaches may improve the accuracy of knowledge extraction and model performance. [20]

Accurate classification of newly diagnosed prostate cancer patients into low- and high-risk at a tertiary care center, where second opinion patients diagnosed outside of the center and patients with complex histories, multiple comorbidities and advanced disease are common is a challenge for any automated data extraction pipeline. Accurately classifying information on bone scan receipt in EHRs is challenging and requires the fusion of heterogeneous data and the development of different data methods.

In this study, we classified prostate cancer patients into risk categories and assessed adherence to guideline recommendations on the need for a bone scan using both structured and unstructured EHR data. We compared the results of an NLP rule-based model and a deep learning model. We measured adherence to both the NCCN and AUA guidelines for avoidance of bone scan for staging in low-risk patients (overuse) and use of a bone scan for staging in high-risk patients (underuse). We demonstrated the utility of gathering multiple data sources captured in diverse formats to assess the efficient and effective use of bone scans for cancer staging among prostate cancer patients.

2. Methods

A graphical outline of our methods to detect the bone scan use with structured and unstructured data from EHRs, can be found in Fig. 1.

2.1. Data source

Patients were identified in a prostate cancer clinical data warehouse, which is described in detail elsewhere. [21] In brief, data were collected from a tertiary-care academic medical center using the Epic EHR system (Epic Systems, Verona, WI) and managed in an EHR-based relational database. Patients were linked to an internal cancer registry and the California Cancer Registry (CCR) to gather additional information on treatments outside the institute, recurrence and survival. This study received the approval from the institute's Institutional Review Board (IRB).

2.2. Study cohort

The study included patients diagnosed with prostate cancer between January 1, 2008 and December 31, 2017. We excluded patients not receiving primary treatment at our medical center and those missing clinical stage, PSA, and Gleason score. PSA is a serum biomarker protein that identifies patients at risk for prostate cancer. For men who have prostate cancer, serum PSA level is associated with prognosis and is used in risk classification. Gleason score is a prognostic grading system that is assigned by a pathologist on prostate cancer tissue samples that is also used in risk classification. Patients were also excluded if they did not have a clinical note in the EHR prior to their primary treatment. Patient and clinical demographics were captured at the time of diagnosis. As guidelines recommend bone scan use after diagnosis and before first treatment, we restricted the data capture procedures to documentation between these dates.

2.3. Risk classification

The NCCN and AUA guidelines classify patients into different groups according to their risk of developing prostate cancer: high risk, intermediate/unfavorable risk, intermediate/favorable risk, and low risk. These classifications are based on clinical tumor stage, PSA value and pre-treatment biopsy Gleason score (Table 1). NCCN guidelines classify patients into several categories: very low, low, favorable intermediate, unfavorable intermediate, high, very high, regional, and metastatic. The categories regional and metastatic are not applicable to this study. We collapsed the NCCN categories into Low- and High-risk groups, since these had historically been used to determine whether a bone scan

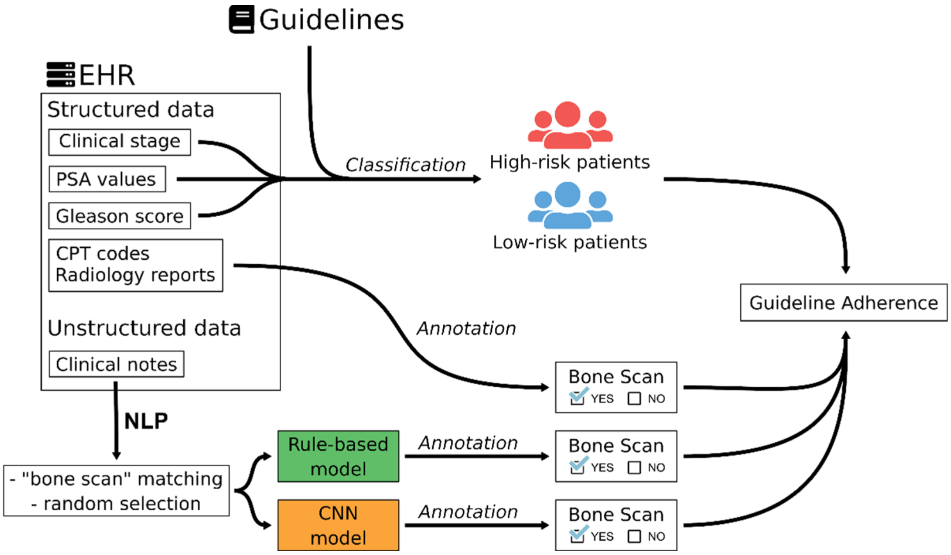

Fig. 1. Illustration of our approach to detect if patients underwent a bone scan.

Table 1
Risk classification groups and inclusion criteria by prostate cancer clinical guidelines.

Guidelines	Risk group	Criteria	Number of patients
NCCN	High risk	Cancer stage T3 or T4 Gleason score ≥ 8 Cancer stage T2 and PSA > 10 ng/mL Cancer stage T1 and PSA > 20 ng/mL	1047
	Low risk	T2 and PSA ≤ 10 and Gleason < 8 T1 and PSA ≤ 20 and Gleason < 8	1168
AUA	High risk	Cancer stage T3 or T4 Gleason Grade Group 4 or Grade Group 5 PSA > 20 ng/mL	989
	Intermediate unfavorable risk	Gleason Grade Group 3 Gleason Grade Group 2 and 10 ≤ PSA < 20 Cancer stage T2B or T2C and Grade Group 2	510
	Intermediate favorable risk	Gleason Grade Group 2 and PSA < 10 Gleason Grade Group 1 and 10 ≤ PSA < 20	559
	Low risk	Cancer stage T1 or T2A and Gleason Grade Group 1 and PSA < 10	469

was recommended. Depending on the patient's risk group, recommendations for bone scan performance are provided. For risk-classification, we assumed information extracted from structured data were accurate and therefore did not perform a manual review, especially because patients with missing data were removed from the cohort. Moreover, the urologists worked very closely with the engineers on hundreds of patient cases to ensure the final risk-classification criteria was accurate.

Criteria for high risk patients are presented in Table 1, which include a combination of overall clinical stage, Gleason score or Gleason grade group, and PSA values. Overall clinical stage was identified in 2 separate structured fields: the institutional cancer registry and the EHRs. When discrepancies occurred, we used the values from the institutional cancer registry as the gold standard. NCCN uses Gleason score and AUA uses Gleason grade group from biopsy. Both variables available in the internal cancer registry and the CCR. When multiple Gleason scores were available, we used the maximum value prior to primary treatment. Finally, PSA values were identified from the laboratory values in the EHR and the CCR. If the patient had a primary treatment, we used the PSA value closest to treatment start date or the PSA value at time of diagnosis. The clinical phenotypes will be available on PheKB.org (see supplementary material 1). [22]

2.4. Detection of bone scan from structured and semi-structured data

Current Procedural Terminology (CPT) codes were used to identify bone scan orders in the EHR structured data: 78300, 78305, 78306, 78315, and 8320. Next, we extracted metadata (considered as semi-structured data) of radiology reports. Each report includes a short description field that indicates the type of radiologic test. A clinician manually selected radiology reports with a description including the expressions “NUC BONE SCAN”, “NM BONE WHOLE BODY” or “NM BONE SCAN”. If a patient had either a CPT code or a radiology report recorded, we considered this evidence that he had received a bone scan.

2.5. Detection of bone scan from unstructured data – Natural Language Processing (NLP) pipelines

Many patients seen at our institution come for a second opinion and therefore receive bone scans outside of our healthcare system. These patients may provide the results from the external bone scan as a paper document or image file of the radiographs that may be recorded by the clinician in the narrative text. We developed a pipeline to extract the information from narrative text which included several types of clinical notes: procedure reports, progress notes, consultation notes, letters, and telephone encounters.

2.6. Data sets

We employed two NLP methods, a rule-based method and a convolutional neural network (CNN) method. For each method, we used the same randomly selected dataset (408 patients) from 5500 patients with 369,764 clinical notes (procedure reports, progress notes, consultation notes, letters, telephone encounters, etc.). We selected randomly one note for each 408 patients containing the word “bone scan” and split these notes in two datasets: a training note-set of 308 notes (76%) and a test note-set of 100 notes (24%). Other terms were considered (i.e. “nuclear study” and “nuclear scan”), however in our corpus of notes these terms referred to other nuclear medicine tests (i.e. not a staging bone scan for prostate cancer) and therefore were not used to filter sentences. For the training note-set, we only included sentences containing the word “bone scan”. For the notes that have more than one sentence with the term “bone scan”, we only selected one of the sentences at random. At the end, our training sentence-set consisted of 308 sentences that were manually annotated; 238 sentences that mention the utilization of a bone scan (positive sentences) and 70 sentences that mention the non-utilization of a bone scan (negative sentences). Finally, to evaluate the accuracy of both models, one clinician manually annotated the test note-set of 100 notes that comprised our gold standard. While we trained the models at the sentence-level, we tested on the note level because to manually annotate hundreds of sentences is resource intensive and from a clinical standpoint, we are concerned with the annotation at the note-level.

For manual annotations, we performed an agreement analysis with a sub part of the training sentence-set. Four clinical researchers annotated a total of 100 sentences [MK, KK, JP, THB]. The Kappa score was 0.89 (see supplementary material 2). Based on the strong agreement between annotators, the remaining 208 sentences in the training set were annotated by a single clinician [KK]. The 100 note test-set was annotated by the research nurse [MF]. The objective of annotation was to identify whether a note had a positive or negative mention of a bone scan.

2.7. Pre-processing

The NLP pipeline first pre-processed each clinical note, which entailed splitting the note into individual sentences, removing capitalization, numbers and punctuation, and excluding words smaller than three letters, except the word “no” and the abbreviation “NM” (Nuclear Medicine). Through this process, a note corresponded to a list of sentences and a sentence corresponded to a list of words. “Bone scan” was the only target key term.

2.8. Rule-based method

The rule-based method applied a set of syntax rules to predict whether a sentence contained information related to a bone scan. The model used the ConText algorithm developed by Chapman et al [23]. ConText is an algorithm derived from the NegEx algorithm to identify negative results in a free text. From regular expressions, it determines whether information in clinical reports are mentioned as negated, hypothetical, historical, or experienced by someone other than the patient. For this study, if bone scan information is negated, hypothetical or historical then we concluded the patient did not receive a bone scan for this note. In addition, if no modifier could be apply to the sentence then, by default, we classified the sentence as negated. We used 90% of the training dataset to build the rules manually and the remaining 10% to validate the model. This iterated process of rule building was used to develop the model.

2.9. Convolutional neural network method

After notes were pre-processed, we used the word2vec method implemented in Gensim [24] to form word embeddings. [25] Word2Vec is a technique to create a vector representing the semantic context of a word for each word in our corpus. If similar words share common contexts in the corpus, then it is assumed they have similar vectors. The word2vec method is self-supervised machine learning method that trains a 2-layer neural network to form word embeddings. Word2vec has two different architectures (skip-gram and Continuous Bag of Words (CBOW)) and two different algorithms (hierarchical softmax and negative sampling). We chose to generate vectors with a dimension of 300. We tried multiple configurations (described in supplementary material 3) and found that for our dataset the best configuration was a combination of the CBOW architecture and the hierarchical softmax algorithm. We also tried different window widths (i.e. the maximum distance between the current and predicted word within a sentence) and we chose a window width of 5.

From the word embeddings, we created a two-dimensional matrix for each sentence where each row corresponded to a word in the sentence and each column to a dimension of the vector. Using this matrix, we applied the convolutional neural network (CNN) method to classify sentences. [26] CNN methods require a uniform size matrix as input. Therefore, we calculated that the maximum sentence size in the notes was 361. If the size of a sentence was smaller than 361, then we completed the sentence with a padding of “0”. Finally, each sentence corresponded to a matrix of 300 × 361.

The model architecture was implemented with the library TensorFlow [27] and was trained on the training data set. We tuned the model using the strategy described by Zhang and Wallace. [28] We used 10 fold cross-validation to validate the model and examined various CNN configuration. The parameter tuning (both word2vec and CNN) was conducted using the training data (with validation splits) only. The tuning strategy and the results are described in the supplementary material 3. The most accurate CNN model consisted of the parameters below:

- filter region size = (3, 4, 5)
- feature maps = 100
- activation function = ReLU
- pooling = 1-max pooling
- dropout rate = 0.6
- l2 norm constraint = 3

2.10. Prediction and evaluation of bone scan utilization

If the model predicted a positive sentence (i.e. the note mentioned utilization of a bone scan) then the entire note was flagged positive. If a patient had at least one positive note between diagnosis and first treatment, we annotated that the patient had received a bone scan. To evaluate the accuracy of both models, the gold standard dataset of 100 notes was used to calculate precision, recall and F-measure.

2.11. Statistical analysis

We performed a statistical analysis of clinical characteristics between the bone scan status for each risk group. The characteristics included the variables age at diagnosis, insurance payor type, ethnicity and race. This analysis consisted of unpaired t-tests for parametric data, between 4 different risk groups (high, unfavorable/intermediate, favorable/intermediate, low) consisted of analysis of variance (ANOVA) for parametric data, whereas the chi-square/Fisher's exact tests were used for categorical variables. All statistical tests were 2-sided with a threshold of $p \leq 0.05$ for statistical significance.

3. Results

3.1. Patients meeting the guideline criteria

From a total 5500 patients; 2215 patients had complete information that allowed risk classification according to the NCCN guidelines and 2527 patients according to the AUA guidelines (Fig. 2). Clinical data for the patients extracted from the EHR are summarized in Table 1. Using the NCCN guidelines, 1047 patients (47%) were high risk while 1168 patients (53%) were classified as low risk. For AUA criteria, 989 patients (39%) were considered as high risk, 510 patients (20%) as unfavorable/intermediate risk, 559 patients (22%) as favorable/intermediate risk and 469 patients (18%) as low risk.

3.2. Patient characteristics

Table 2 presents patient demographics stratified by level of risk for the NCCN guidelines (Table 2A) and the AUA guidelines (Table 2B) and by bone scan examination (predictions of the CNN model). Overall, the patient demographics did not differ significantly between the patients who underwent a bone scan compared to those that did not within each risk group. There were statistically significant differences in age, with older patients less likely to receive a bone scan for high risk (68.72 vs 67.05, $p = 0.004$) and more likely to receive a bone scan for low risk cancer (63.38 vs 64.43, $p = 0.054$), although these differences were small and not likely to be not clinically significant.

3.3. Evaluation of NLP models

Table 3 shows the accuracy of the rule-based and CNN models as a function of precision, recall and F-measure values based on 100 manually annotated notes. The rule-based model showed high precision (0.924) but a lower recall (0.871), indicating that the model missed many notes mentioning bone scan utilization, but was rarely wrong regarding positive identification of bone scan performance. The results are different for the CNN model, where precision was not as high (0.882), but recall was very high (0.957).

The predictions of the two models are summarized in Fig. 3, where each node represents a note. Nodes that are black correspond to notes where bone scan utilization was mentioned, whereas white nodes represent notes that state that the patient has not received a bone scan. The orange and green zones correspond to the predictions of the bone scan receipt according to the 2 models.

The combination of the two methods improved the accuracy of information extraction. In Table 3, four possible models are presented that differentially harmonize precision and recall to adjust model accuracy. It was possible to combine the predictions of the two models. For example, the number of false positives could be minimized by using the intersection of notes with positive annotations by both methods (model 3). This approach increased the precision score (0.968) at the

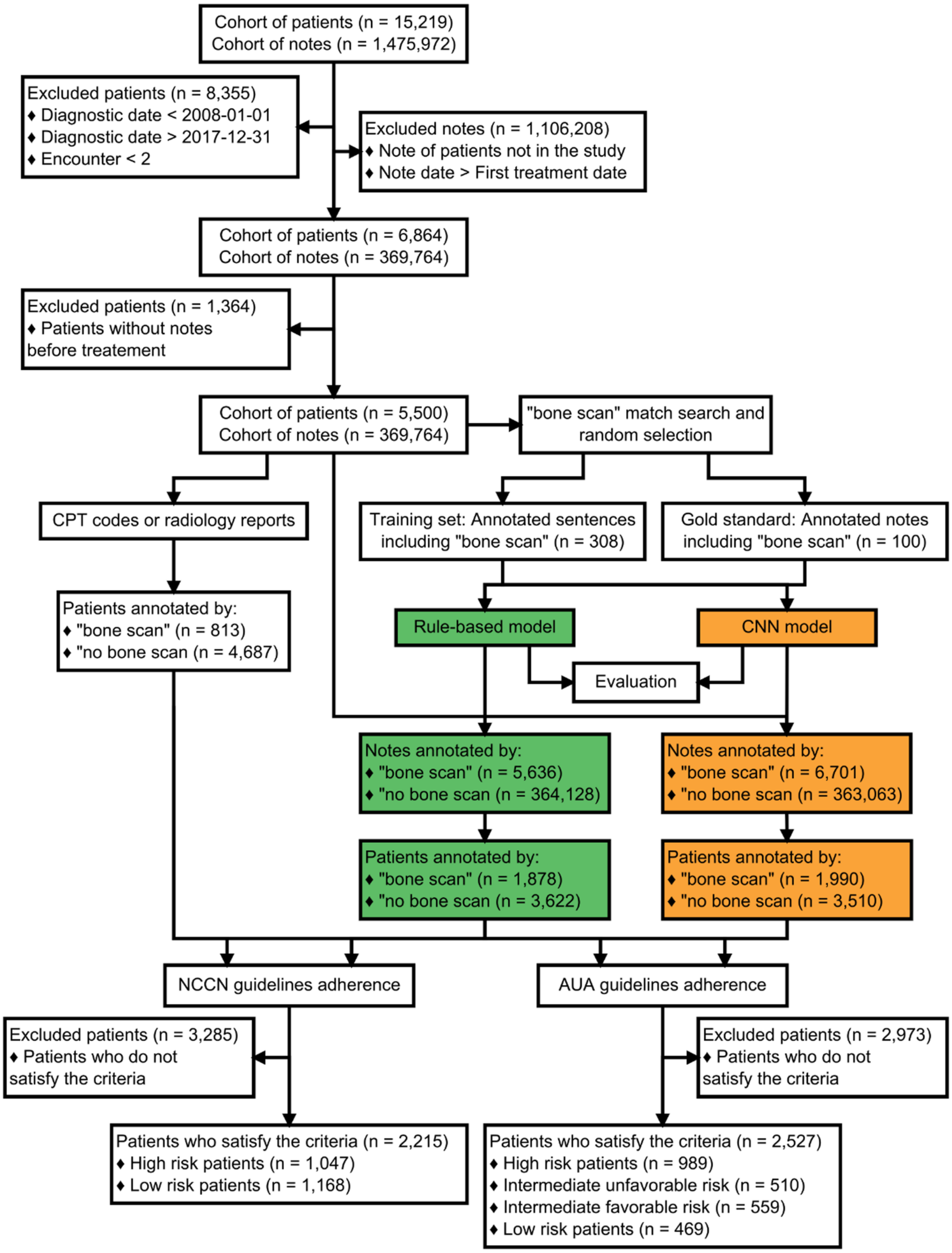
Cohort of patients (n = 15,219)
Cohort of notes (n = 1,475,972)
Excluded patients (n = 8,355)
♦ Diagnostic date < 2008-01-01
♦ Diagnostic date > 2017-12-31
♦ Encounter < 2
Excluded notes (n = 1,106,208)
♦ Note of patients not in the study
♦ Note date > First treatment date
Cohort of patients (n = 6,864)
Cohort of notes (n = 369,764)
Excluded patients (n = 1,364)
♦ Patients without notes before treatement
Cohort of patients (n = 5,500)
Cohort of notes (n = 369,764)
"bone scan" match search and random selection
CPT codes or radiology reports
Training set: Annotated sentences including "bone scan" (n = 308)
Gold standard: Annotated notes including "bone scan" (n = 100)
Patients annotated by:
♦ "bone scan" (n = 813)
♦ "no bone scan (n = 4,687)
Rule-based model
CNN model
Evaluation
Notes annotated by:
♦ "bone scan" (n = 5,636)
♦ "no bone scan (n = 364,128)
Notes annotated by:
♦ "bone scan" (n = 6,701)
♦ "no bone scan (n = 363,063)
Patients annotated by:
♦ "bone scan" (n = 1,878)
♦ "no bone scan (n = 3,622)
Patients annotated by:
♦ "bone scan" (n = 1,990)
♦ "no bone scan (n = 3,510)
NCCN guidelines adherence
AUA guidelines adherence
Excluded patients (n = 3,285)
♦ Patients who do not satisfy the criteria
Excluded patients (n = 2,973)
♦ Patients who do not satisfy the criteria
Patients who satisfy the criteria (n = 2,215)
♦ High risk patients (n = 1,047)
♦ Low risk patients (n = 1,168)
Patients who satisfy the criteria (n = 2,527)
♦ High risk patients (n = 989)
♦ Intermediate unfavorable risk (n = 510)
♦ Intermediate favorable risk (n = 559)
♦ Low risk patients (n = 469)

Fig. 2. Flowchart to select the final cohort, to classify the patients and to detect if patients underwent a bone scan.

Table 2A
Demographic data of patients in function of NCCN guidelines and predictions of CNN model.

Patient characteristics		NCCN risk group					
		High risk (n = 1047)			Low risk (n = 1168)		
		No BS	BS	p	No BS	BS	p
Total, n (%)		354 (33.8)	693 (66.2)		945 (80.9)	223 (19.1)	
Age at diagnosis (years), Ẋ ± sd		68.72 ± 9.2	67.05 ± 8.69	0.004	63.38 ± 7.4	64.43 ± 7.2	0.054
Insurance Payor Type, n (%)	Private	102 (29.4)	245 (70.6)	0.095	448 (84.1)	85 (15.9)	0.020
	Medicare	211 (35.8)	378 (64.2)		422 (77.8)	121 (22.2)	
	Medicaid	14 (28.0)	36 (72.0)		33 (86.8)	5 (13.2)	
Ethnicity, n (%)	Non-Hispanic	323 (34.0)	628 (66.0)	0.227	869 (80.8)	206 (19.2)	0.939
	Hispanic	24 (27.6)	63 (72.4)		69 (81.2)	16 (18.8)	
Race, n (%)	Asian	36 (23.8)	115 (76.2)	0.019	91 (83.5)	18 (16.5)	0.590
	White	261 (35.5)	475 (64.5)		732 (80.2)	181 (19.8)	
	Others	25 (36.8)	43 (63.2)		42 (84.0)	8 (16.0)	

Table 2B
Demographic data of patients in function of AUA guidelines and predictions of CNN model.

Patient characteristics		AUA risk group											
		High risk			Intermediate unfavorable risk			Intermediate favorable risk			Low risk		
		(n = 989)			(n = 510)			(n = 559)			(n = 469)		
		No BS	BS	p	No BS	BS	p	No BS	BS	p	No BS	BS	p
Total, n (%)		267 (27.0)	722 (73.0)		310 (60.8)	200 (39.2)		433 (77.5)	126 (22.5)		424 (90.4)	45 (9.6)	
Age at diagnosis (years), $\bar{X}$ ± sd		69.74 ± 9.78	68.97 ± 9.45	0.262	67.04 ± 7.78	65.66 ± 6.84	0.040	63.94 ± 7.52	63.01 ± 7.75	0.222	62.46 ± 7.43	63.31 ± 6.8	0.459
Insurance Payor Type, n (%)	Private	68 (23.9)	216 (76.1)	0.489	98 (57.3)	73 (42.7)	0.578	206 (77.7)	59 (22.3)	0.939	219 (92.4)	18 (7.6)	0.164
	Medicare	163 (27.7)	426 (72.3)		184 (62.2)	112 (37.8)		194 (76.7)	59 (23.3)		174 (88.3)	23 (11.7)	
	Medicaid	14 (25.0)	42 (75.0)		10 (62.5)	6 (37.5)		12 (75.0)	4 (25.0)		14 (100.0)	0 (0.0)	
Ethnicity, n (%)	Non-Hispanic	241 (27.0)	652 (73.0)	0.355	291 (61.0)	186 (39.0)	0.913	395 (77.5)	115 (22.5)	0.771	387 (90.8)	39 (9.2)	0.454
	Hispanic	19 (22.4)	66 (77.6)		18 (60.0)	12 (40.0)		34 (75.6)	11 (24.4)		34 (87.2)	5 (12.8)	
Race, n (%)	Asian	26 (18.2)	117 (81.8)	0.028	41 (66.1)	21 (33.9)	0.335	44 (80.0)	11 (20.0)	0.848	35 (92.1)	3 (7.9)	0.900
	White	196 (29.0)	479 (71.0)		235 (59.8)	158 (40.2)		333 (76.6)	102 (23.4)		323 (90.2)	35 (9.8)	
	Others	20 (29.0)	49 (71.0)		18 (72.0)	7 (28.0)		17 (77.3)	5 (22.7)		23 (92.0)	2 (8.0)	

Table 3
Evaluation of NLP models in function of 100 manually annotated notes.

	Model 1	Model 2	Model 3	Model 4
	Rule-based	CNN	Rule-based and CNN	Rule-based or CNN
Precision	0.924	0.882	0.968	0.850
Recall	0.871	0.957	0.857	0.971
F-measure	0.897	0.918	0.909	0.907

expense of decreasing the recall score (0.857). False negatives could be minimized by selecting the union of patients of positive annotations by both methods (model 4), producing high recall (0.971) but low precision (0.85).

The 5500 patients included in our study had 369,764 associated notes. These notes were composed of a total of 17,101,187 sentences, including 14,090 sentences with the word “bone scan”. The CNN model predicted 6701 positive notes from the 369,764 notes and the rule-based model predicted 5636 positives notes. The intersection of model predictions (model 3) was 5326 positive notes, while the union of model predictions (model 4) included 7011 positive notes.

3.4. Guideline adherence

To measure guideline adherence, we chose to use the CNN model because it had the highest F-measure (0.918) compared to the rule-based model, and therefore the best compromise between precision and recall. Using structured and semi-structured data, we determined that only 813 patients received a bone scan (15%). However, an additional 1270 patients (23%) were annotated when we used the CNN model. Fig. 4 summarizes the use of bone scan according to the NCCN and AUA guidelines, where each bar corresponds to the percentage of patients who received a bone scan. Bone scans were used at modestly high rates in high-risk patients (73%), while only 10% of low-risk patients received a bone scan. When intermediate risk patients were substratified into unfavorable risk and favorable risk according to the AUA guidelines, 39% and 23% underwent bone scan, respectively.

4. Discussion

We developed a pipeline using heterogeneous EHR data to assess guideline adherence (the over- and under-use) of radionuclide bone scans in newly diagnosed prostate cancer patients for staging prior to treatment. To measure adherence, we developed electronic phenotypes to classify patients into different clinical risk categories according to two different guidelines because each clinical risk category has a different bone scan recommendation. Assessment of bone scan documentation required the transformation of heterogenous data to knowledge using NLP technologies, with CNN models outperforming a rule-based approach. Our work also provides a model for the demonstrates the use of orthogonal NLP methods to adjust model precision for individual use cases, allowing to titrate for higher precision to ensure all high-risk patients needing a bone scan are identified, or for higher recall to measure guideline adherence. For assessment of adherence to the bone scan quality metric, it is critical to avoid false positives (label a high-risk patient as ‘bone scan performed’ if he did not receive one), therefore models tuned to the highest precision would minimize this risk and lower the number of false positives. Integrating this information at point of care will be essential to ensure both patients and clinics have evidence necessary to guide bone scan use and treatment pathways.

Bone scans provide information on cancers that have metastasized into skeletal structures. Pre-treatment metastases are important to identify, as stage dictates the appropriate treatment and provides prognosis for a patient. Bone scans were documented in diffuse sites in the EHRs: as structured data (i.e. CPT codes), semi-structured data (i.e.

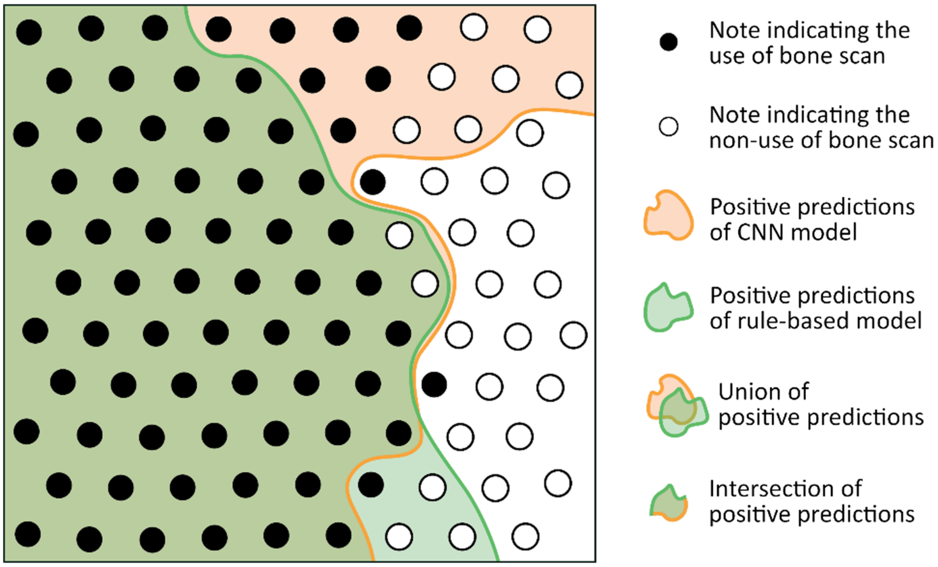
Note indicating the use of bone scan
Note indicating the non-use of bone scan
Positive predictions of CNN model
Positive predictions of rule-based model
Union of positive predictions
Intersection of positive predictions

Fig. 3. Evaluation of NLP model predictions in 100 manual annotated notes. Each node represents a note. If a note mentions that a bone scan was performed, the node is black. If a note mentions that the patient had not received a bone scan or that a bone scan is planned then the note is white.

radiology reports), and unstructured text (clinical narratives). To identify and extract this information, we used both rule-based and machine learning methods. Rule-based models are known to be very conservative because the rules are built manually and cannot cover all possible scenarios. Therefore, this approach displays high precision but low recall, is limited by the variability of documentation, and can only be improved with additional rules. On the other hand, the CNN model is more flexible because each word is represented by a vector and similar words have similar vectors. Therefore, if the neural network learns to classify a sentence in a category, then subsequent sentences containing similar words will have a higher probability of being classified in the same category. CNN models have lower precision but higher recall compared to rule-based approaches. Therefore, we used a combination of the model predictions to balance the precision and the recall based on our particular question. Our results indicate that by using different iterations of the two NLP models, we can toggle between high precision and high recall depending on the research question or clinical need. As government and the health care industry begin to incorporate real world evidence from EHRs into regulatory and evaluation purposes, these different methods ensure the high accuracy and flexibility to adjust output to fit regulatory or clinical needs. [29]

We find significant under-estimation of bone scan documentation when using structured data alone. Our data suggest that advanced technologies to leverage unstructured data buried in EHRs are needed to accurately assess certain electronic phenotypes, such as those related to treatment pathways or risk categories, as we have shown in other work. [30,31] The limitations of EHR structured data regarding missingness and accuracy is concerning, especially as many studies focus only on these data for clinical phenotyping. Therefore, the use of advanced technologies, such as neural networks, on unstructured clinical narrative text will be critical for improving model accuracy, particularly when assessing guideline adherence where payment incentives and penalties might be relevant.

To classify patients into the different risk categories, clinical information was needed from multiple data sources at specific time periods during the care pathway. Such clinical phenotyping is a

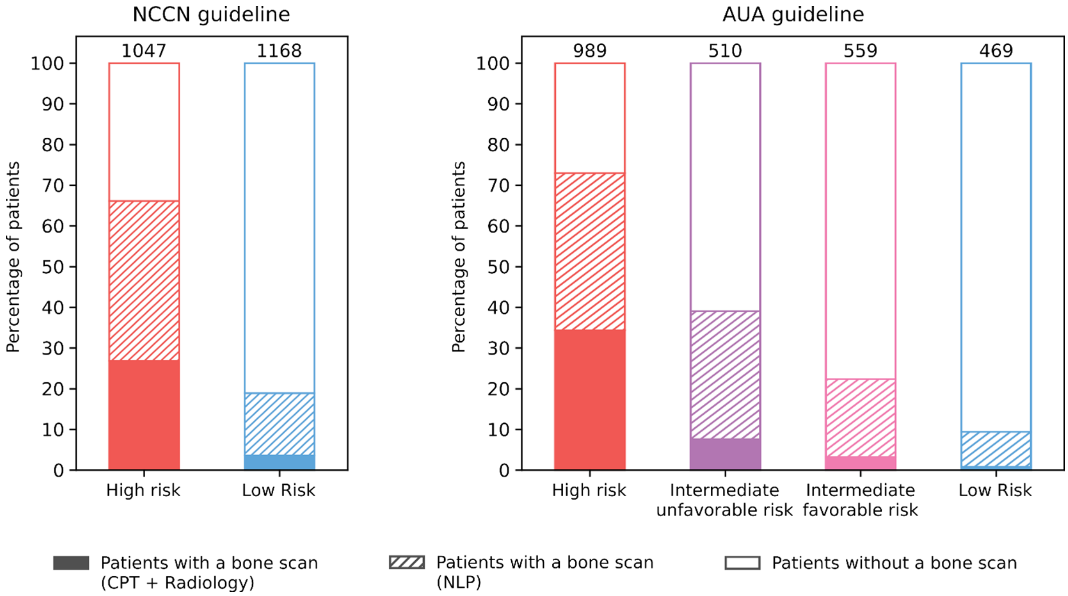

Fig. 4. Guideline adherence. Percentage of patients undergoing a bone scan stratified by risk group according to the NCCN and AUA guidelines.

fundamental task necessary to use EHRs for secondary research, which may include both rule-based and machine learning approaches. [32] In this study, we synthesized the granular digital data down to the patient level. This included diagnostic information (e.g. PSA levels, Gleason score) and clinical prognostic factors (e.g. summary stage), which were collected from multiple sources in the clinical data warehouse. Using this information, we classified patients into the categories necessary to assess guideline adherence: high-risk, intermediate-risk, and low-risk. Such classifications are essential to assess prognosis, treatment pathways, and quality of care.

Using our methods, we found the majority of high-risk patients had received a bone scan while only 10% of low-risk patients had one, which is in accordance with both the AUA and NCCN guidelines. Since we could link directly to patient demographics in the EHR, we were able to determine that the over-use or under-use of bone scans did not differ by patient characteristics. Use of bone scans in intermediate risk patients is controversial, since only a small fraction of these patients will harbor metastatic disease detectable on a bone scan. [33] This controversy is reflected in the relatively low bone scan rates of 30–40% in our practice that includes many providers. Reducing the over-use of bone scans in low risk patients has been identified in the Physician Quality Reporting System, both to cut down unnecessary health care expenditures, and to decrease unnecessary radiation exposure. [34] Bone scan utilization is also a quality metric and is used by the Center of Medicare and Medicaid services an subject to payment penalties. [35] The methods we have developed could be used for quality metric capture and reporting, both at the level of the individual clinician level and at the department, practice and hospital level. Direct feedback on inappropriate use of imaging in prostate cancer has been shown to favorably alter physician behavior. [36]

It is important to understand why some high-risk patients did not receive a bone scan while other low-risk patients did receive one. Often, guideline recommendations may not be available at point of care and therefore patients in need of a bone scan may be missed. We found that this was frequently the case when patients were classified as high risk based only a single variable, i.e. PSA > 20 or Gleason Grade > 4. However, these variables can also be erroneous recorded in registry data [37] and therefore the clinician may be providing care consistent with guidelines. For low-risk patients, a complaint of back pain could signal bone metastases and guidelines state this is an indication for a bone scan. The presence of symptoms, such as bone pain, is often not recorded in the EHR. Therefore, while the recommendations from the AUA and NCCN assist in clinical decisions regarding bone scans for cancer staging, individual patients may present with additional criteria that may signal alternative care pathways.

Important differences between the predictions of the CNN model and the rule-based model were identified. For the rule-based model, by default a sentence was negated if no rule could be applied. This decision was made because in the training set, many sentences included the word “bone scan” but they were describing guideline recommendations and were not associated with performing a bone scan for the patient. This property is not true for the CNN model therefore with the above example, the rule-based model correctly predicts the sentence as negative while the CNN model could incorrectly predict it as positive. On the other hand, in the validation set we found some sentences mentioning the use of bone scan for patients for which no rule existed because these sentences were not present in our training set. However, the CNN model includes the word embeddings generated by word2vec, where words with a similar semantic have similar vectors. In some cases, this property can allow the CNN model to correctly label some sentences with unknown formulations. These important differences suggest that the CNN model might be a better solution for a decision-support infrastructure because it is possible to create a feedback system where the model can learn over time, while the rule based model would need to have continuous manual rule building.

Our work has several limitations that should be mentioned. First, our algorithms were constructed on an EHR from one institute. However, the records encompass diverse providers (physicians, nurse practitioners, physician’s assistants) and several practice settings (surgical, medical radiation oncology, and primary care). In addition, we have made our algorithms publicly available for validation elsewhere, since privacy protections prohibit institutions from allowing us to test our algorithms in their EHR directly. Second, even after pulling information from multiple data sources, only 40% of our eligible population had complete data capture necessary for classification. A significant portion of the key variables (e.g. Gleason score, PSA) needed to classify cancer patients to appropriate risk categories were missing from the CCR and EHRs, as we have previously reported. [21] This is often the case for academic, tertiary care medical centers that have a large proportion of patients seeking second opinions, where patients do not receive initial biopsies or imaging at the tertiary center and therefore the results of these tests may not be recorded in the EHR system. Interoperability would mitigate this issue, however healthcare systems still struggle with data sharing and care coordination. [38] Third, the imbalance in our training set might affect the results, as there were many more positive than negative sentences. Buda et al concluded that the imbalance effect has a significant impact on prediction. [39] However, the imbalance ratios tried in their study (from 0 to 50) were higher than those from our study. A future strategy could use oversampling from the underrepresented sample to address the imbalance in the dataset. Fourth, radiology reports used in this study were semi-structured. After consultation with a radiologist and a manual review of the expressions, only “NUC BONE SCAN“, ”NM BONE WHOLE BODY“ and ”NM BONE SCAN” were identified as relevant to this study. Other institutes wishing to replicate our study may use additional expressions and terms. Fifth, our training dataset was limited to 300 sentences, which resulted in missed rules. However, there is always a balance between resources and for this project, we limited the training dataset. Future iterations of this pipeline could expand the training dataset, however it is unclear how many additional sentences would be required to significant improve the performance of the model. Finally, some of our data, such as the PSA values, were extracted from the registries (the institutional Cancer Center Registry and the California Cancer Registry), and we have reported previously that they are subject to data entry errors. These errors could bias our results. Fortunately, the number of these errors are relatively small in the population and infrequently affect risk group classification in < 5% of cases. [40]

5. Conclusion

We have developed a method for prostate cancer patient risk stratification and extraction of bone scan performance using 2 NLP models for monitoring adherence to quality metrics by combining structural and non-structural data from EHRs. The model based on a convolutional neural network obtained better results than the rule-based model; however, a combination of the two models to optimize performance to suit individual use cases can be used to optimize the quality of the annotations. While adherence with guidelines in our practice was very good, documentation of adherence allows opportunities for quality improvement at an individual or practice level. Our method could serve as the basis of a decision-support algorithm to provide decision support for practitioners.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA183962. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Declaration of interests

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.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jbi.2019.103184.

References

[1] M.M. Center, A. Jemal, J. Lortet-Tieulent, et al., International variation in prostate cancer incidence and mortality rates, Eur. Urol. 61 (2012) 1079–1092.

[2] M.A. Dall'Era, P.C. Albertsen, C. Bangma, et al., Active surveillance for prostate cancer: a systematic review of the literature, Eur. Urol. 62 (2012) 976–983.

[3] A.J. Chang, K.A. Autio, M. Roach Iii, et al., High-risk prostate cancer—classification and therapy, Nat. Rev.Clin. Oncol. 11 (2014) 308–323.

[4] A.V. D'Amico, R. Whittington, S.B. Malkowicz, et al., Biochemical outcome after radical prostatectomy, external beam radiation therapy, or interstitial radiation therapy for clinically localized prostate cancer, JAMA 280 (1998) 969–974.

[5] A.D. Falchook, R.G. Salloum, L.H. Hendrix, et al., Use of bone scan during initial prostate cancer workup, downstream procedures, and associated medicare costs, Int. J. Radiat. Oncol. 89 (2014) 243–248.

[6] D.W. Blayney, K. McNiff, D. Hanauer, et al., Implementation of the quality oncology practice initiative at a university comprehensive cancer center, J. Clin. Oncol. 27 (2009) 3802–3807.

[7] J.L. Mohler, A.J. Armstrong, R.R. Bahnson, et al., Prostate Cancer, Version 1.2016, J. Natl. Compr. Canc. Netw. 14 (2016) 19–30.

[8] H.B. Carter, P.C. Albertsen, M.J. Barry, et al., Early detection of prostate cancer: AUA guideline, J. Urol. 190 (2013) 419–426.

[9] J.M. Albert, P. Das, Quality indicators in radiation oncology, Int. J. Radiat. Oncol. Biol. Phys. 85 (2013) 904–911, https://doi.org/10.1016/j.ijrobp.2012.08.038.

[10] C.P. Filson, Quality of care and economic considerations of active surveillance of men with prostate cancer, Transl. Androl. Urol. 7 (2018) 203–213, https://doi.org/10.21037/tau.2017.08.08.

[11] A.D. Falchook, L.H. Hendrix, R.C. Chen, Guideline-discordant use of imaging during work-up of newly diagnosed prostate cancer, J. Oncol. Pract. 11 (2015) e239–e246.

[12] W.-W. Yim, A.J. Wheeler, C. Curtin, et al., Secondary use of electronic medical records for clinical research: challenges and opportunities, Converg. Sci. Phys. Oncol. 4 (2018) 014001.

[13] J. Adler-Milstein, C.M. DesRoches, P. Kralovec, et al., Electronic health record adoption in US hospitals: progress continues, but challenges persist, Health Aff. (Millwood) 34 (2015) 2174–2180.

[14] S. Velupillai, H. Suominen, M. Liakata, et al., Using clinical Natural Language Processing for health outcomes research: overview and actionable suggestions for future advances, J. Biomed. Inform. 88 (2018) 11–19, https://doi.org/10.1016/j.jbi.2018.10.005.

[15] C.A. Bejan, J. Angiolillo, D. Conway, et al., Mining 100 million notes to find homelessness and adverse childhood experiences: 2 case studies of rare and severe social determinants of health in electronic health records, J. Am. Med. Inform. Assoc. 25 (2018) 61–71.

[16] K.B. Wagholikar, K.L. MacLaughlin, M.R. Henry, et al., Clinical decision support with automated text processing for cervical cancer screening, J. Am. Med. Inform. Assoc. 19 (2012) 833–839.

[17] L.T.E. Cheng, J. Zheng, G.K. Savova, et al., Discerning tumor status from unstructured MRI reports—completeness of information in existing reports and utility of automated natural language processing, J. Digit. Imaging 23 (2010) 119–132.

[18] B. Percha, Y. Zhang, S. Bozkurt, et al., Expanding a radiology lexicon using contextual patterns in radiology reports, J. Am. Med. Inform. Assoc. 25 (2018) 679–685.

[19] C. Xiao, E. Choi, J. Sun, Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 25 (2018) 1419–1428.

[20] L. Deleger, H. Brodzinski, H. Zhai, et al., Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department, J. Am. Med. Inform. Assoc. 20 (2013) e212–e220.

[21] M.G. Seneviratne, T. Seto, D.W. Blayney, et al., Architecture and implementation of a clinical research data warehouse for prostate cancer, EGEMs Gener. Evid Methods Improve Patient Outcomes 6 (2018) 13.

[22] J.C. Kirby, P. Speltz, L.V. Rasmussen, et al., PheKB: a catalog and workflow for creating electronic phenotype algorithms for transportability, J. Am. Med. Inform. Assoc. JAMIA 23 (2016) 1046–1052.

[23] B.E. Chapman, S. Lee, H.P. Kang, et al., Document-level classification of CT pulmonary angiography reports based on an extension of the ConText algorithm, J. Biomed. Inform. 44 (2011) 728–737.

[24] R. Řehůřek, P. Sojka, Software framework for topic modelling with large corpora, Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA, 2010, pp. 45–50.

[25] T. Mikolov, K. Chen, G. Corrado, et al., Efficient Estimation of Word Representations in Vector Space, ArXiv13013781 Cs Published Online First: 16 January 2013, http://arxiv.org/abs/1301.3781 (accessed 16 Oct 2018).

[26] Y. Kim, Convolutional Neural Networks for Sentence Classification, ArXiv14085882 Cs Published Online First: 25 August 2014, http://arxiv.org/abs/1408.5882 (accessed 22 Aug 2018).

[27] Martín Abadi, Ashish Agarwal, Paul Barham, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015. https://www.tensorflow.org/.

[28] Y. Zhang, B. Wallace, A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification, ArXiv151003820 Cs Published Online First: 13 October 2015, http://arxiv.org/abs/1510.03820 (accessed 22 Aug 2018).

[29] 21st Century Cures Act. H.R. 34, 114th Congress, 2016, https://www.congress.gov/bill/114th-congress/house-bill/34/text (accessed 9 Nov 2018).

[30] W.-Q. Wei, J.C. Denny, Extracting research-quality phenotypes from electronic health records to support precision medicine, Genome Med. 7 (2015) 41, https://doi.org/10.1186/s13073-015-0166-y.

[31] V. Agarwal, T. Podchiyska, J.M. Banda, et al., Learning statistical models of phenotypes using noisy labeled training data, J. Am. Med. Inform. Assoc. 23 (2016) 1166–1173, https://doi.org/10.1093/jamia/ocw028.

[32] J.M. Banda, M. Seneviratne, T. Hernandez-Boussard, et al., Advances in electronic phenotyping: from rule-based definitions to machine learning models, Annu. Rev. Biomed. Data Sci. 1 (2018) 53–68, https://doi.org/10.1146/annurev-biodatasci-080917-013315.

[33] G.V. KandaSwamy, A. Bennett, K. Narahari, et al., Establishing the pathways and indications for performing isotope bone scans in newly diagnosed intermediate-risk localised prostate cancer - results from a large contemporaneous cohort, BJU Int. 120 (2017) E59–E63, https://doi.org/10.1111/bju.13850.

[34] N. Anumula, P.C. Sanelli, Physician quality reporting system, Am. J. Neuroradiol. 32 (2011) 2000–2001, https://doi.org/10.3174/ajnr.A2912.

[35] D. Gori, R. Dulal, D.W. Blayney, et al., Utilization of prostate cancer quality metrics for research and quality improvement: a structured review, Jt. Commun. J. Qual. Patient. Saf. (2018), https://doi.org/10.1016/j.jcjq.2018.06.004.

[36] A.B. Rutledge, N. McLeod, N. Mehan, et al., A clinician-centred programme for behaviour change in the optimal use of staging investigations for newly diagnosed prostate cancer, BJU Int. 121 (Suppl 3) (2018) 22–27, https://doi.org/10.1111/bju.14144.

[37] D.P. Guo, I.-C. Thomas, H.R. Mittakanti, et al., The research implications of prostate specific antigen registry errors: data from the veterans health administration, J. Urol. (2018), https://doi.org/10.1016/j.juro.2018.03.127.

[38] J.M. Madden, M.D. Lakoma, D. Rusinak, et al., Missing clinical and behavioral health data in a large electronic health record (EHR) system, J. Am. Med. Inform. Assoc. JAMIA 23 (2016) 1143–1149, https://doi.org/10.1093/jamia/ocw021.

[39] M. Buda, A. Maki, M.A. Mazurowski, A systematic study of the class imbalance problem in convolutional neural networks, Neural Netw. 106 (2018) 249–259, https://doi.org/10.1016/j.neunet.2018.07.011.

[40] H.R. Mittakanti, I.-C. Thomas, J.B. Shelton, et al., Accuracy of prostate-specific antigen values in prostate cancer registries, J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 34 (2016) 3586–3587, https://doi.org/10.1200/JCO.2016.68.9216.