*JAMIA Open*, 2(1), 2019, 150–159
doi: 10.1093/jamiaopen/ooy057
Advance Access Publication Date: 4 January 2019
Research and Applications

Research and Applications

# Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment

**Imon Banerjee,[1] Kevin Li,[2] Martin Seneviratne,[1,3] Michelle Ferrari,[4] Tina Seto,[5] James D. Brooks,[4] Daniel L. Rubin[1,6,7,*] and Tina Hernandez-Boussard[1,7,8,*]**

[1]Department of Biomedical Data Science, Stanford University School of Medicine, Medical School Office Building (MSOB), 1265 Welch Road, Stanford, California 94305-5479, USA, [2]Stanford University School of Medicine, 291 Campus Drive, Stanford, California 94305-5479, USA, [3]Department of Biomedical Informatics, Stanford University School of Medicine, Medical School Office Building (MSOB), 1265 Welch Road, Stanford, California 94305-5479, USA, [4]Department of Urology - Divisions, Stanford University School of Medicine, 875 Blake Wilbur, Stanford, California 94305-5479, USA, [5]IRT Research Technology, Stanford University School of Medicine, Stanford, California 94305-5479, USA, [6]Department of Radiology, Stanford University School of Medicine, Stanford, California 94305-5479, USA, [7]Department of Medicine (Biomedical Informatics), Stanford University School of Medicine, Medical School Office Building (MSOB), 1265 Welch Road, Stanford, California 94305-5479, USA and [8]Department of Surgery, Stanford University School of Medicine, 300 Pasteur Drive Stanford, California 94305-2200, USA

*The last two authors contributed equally.

Corresponding Author: Imon Banerjee, Department of Biomedical Data Science, Stanford University School of Medicine, Medical School Office Building (MSOB), 1265 Welch Road, Stanford, CA 94305-5479, USA (imonb@stanford.edu)

Received 14 August 2018; Revised 14 November 2018; Editorial Decision 16 November 2018; Accepted 28 November 2018

## ABSTRACT

**Background:** The population-based assessment of patient-centered outcomes (PCOs) has been limited by the efficient and accurate collection of these data. Natural language processing (NLP) pipelines can determine whether a clinical note within an electronic medical record contains evidence on these data. We present and demonstrate the accuracy of an NLP pipeline that targets to assess the presence, absence, or risk discussion of two important PCOs following prostate cancer treatment: urinary incontinence (UI) and bowel dysfunction (BD).

**Methods:** We propose a weakly supervised NLP approach which annotates electronic medical record clinical notes without requiring manual chart review. A weighted function of neural word embedding was used to create a sentence-level vector representation of relevant expressions extracted from the clinical notes. Sentence vectors were used as input for a multinomial logistic model, with output being either presence, absence or risk discussion of UI/BD. The classifier was trained based on automated sentence annotation depending only on domain-specific dictionaries (weak supervision).

**Results:** The model achieved an average F1 score of 0.86 for the sentence-level, three-tier classification task (presence/absence/risk) in both UI and BD. The model also outperformed a pre-existing rule-based model for note-level annotation of UI with significant margin.

**Conclusions:** We demonstrate a machine learning method to categorize clinical notes based on important PCOs that trains a classifier on sentence vector representations labeled with a domain-specific dictionary, which eliminates the need for manual engineering of linguistic rules or manual chart review for extracting the PCOs. The weakly supervised NLP pipeline showed promising sensitivity and specificity for identifying important PCOs in unstructured clinical text notes compared to rule-based algorithms.

**Trial registration:** This is a chart review study and approved by Institutional Review Board (IRB).

**Key words**: natural language processing, patient-centered outcomes, prostate cancer, neural word embedding, text mining

## INTRODUCTION

Prostate cancer is the most common noncutaneous malignancy in men, accounting for 19% of new cancer diagnoses in the United States in 2017.[1] Multiple treatment modalities exist, including surgery and radiotherapy.[2] These treatments are known to be associated with treatment-related side effects that can alter a patient's quality of life, such as sexual, urinary, and bowel dysfunction (BD).[3] These outcomes are not detectable by a labaratory or diagnostic test, but rather through patient communication and they are often referred to as patient-centered outcomes (PCOs).[4] Therefore, the data are typically captured as free text in clinical narrative documents or through patient surveys, if at all,[5,6] both which are labor-intensive and subject to biases. However, with relative 5-year survival in low-risk localized prostate cancer now above 99%,[7] these treatment-related side effects have emerged as an important discriminator in prostate cancer care management and treatment decisions and more evidence-based research on these outcomes can assist both patients and clinicians to make informed decisions about treatment pathways, promoting value-based care.[8,9] Furthermore, the assessment and documentation of these outcomes are proposed quality metrics for prostate cancer care and under consideration for value-based payment modifiers under healthcare reform.[9,10] Therefore, efforts to efficiently and accurately assess these outcomes align with the principles of value-based care and forms part of a growing national research agenda around patient-centered care.

Computerized natural language processing (NLP) techniques can potentially be a solution for parsing millions of free text clinical narratives stored in hospital repositories, extracting PCOs, and converting them into a structured representation, including both supervised machine-learned and rule-based strategies. Such strategies have already been applied to a range of clinical notes, including progress notes and radiology and pathology reports to extract relevant clinical information in structured format.[11] Supervised machine learning for automatic extraction of information from clinical narratives are common.[12–15] In the prostate cancer domain, NLP offers an opportunity to extract treatment-related side effects on a large-scale from historical notes, which may help train models to automatically predict these outcomes for future patients. Developing such an NLP pipeline would enable secondary analyses on these data and help to provide valuable population-based evidence on these important outcomes. Previous NLP studies in prostate cancer applied rule-based strategies to classify whether a clinical note contained evidence of urinary incontinence (UI), mapping tokens in the note against a dictionary of related terms with a negation detection system, yielding reasonable precision and recall compared to manual chart review.[16,17] However, building supervised systems requires large amounts of annotated data, which is tedious and time-consuming to produce and a core limitation of such systems is their generalizability to other locations and settings.

Recent advances in NLP techniques can be leveraged for the automatic interpretation of free-text narratives by exploiting distributional semantics to provide adequate generalizability by addressing linguistic variability.[18,19] Yet such techniques need a small subset of annotated data for training supervised classifiers when manual annotations are a major limitation. A *weakly supervised approach* is a promising technique for various NLP tasks aimed to minimize human effort by creating training data heuristically from the corpus content or exploiting the pre-existing domain knowledge. Following this idea, we propose a weakly supervised machine learning method for extracting treatment-related side effects following prostate cancer therapy from multiple types of clinical notes.

We extend previous studies both clinically and methodologically, with the objective to extract both treatment-related: UI and BD from a range of clinical notes without considering manually engineered classification rules or large-scale manual annotations. For machine learning, the method exploits two sources of pre-existing medical knowledge: (1) domain-specific dictionaries that have been previously developed for implementing a rule-based information extraction systems;[17] and (2) publicly available CLEVER terminology (https://github.com/stamang/CLEVER/blob/master/res/dicts/base/clever_base_terminology.txt) that represents a vocabulary of terms that often present within clinical narratives. A weighted neural word embedding is used to generate sentence-level vectors where term weights are computed using term frequency and inverse document frequency (TF-idf) scoring mechanism, with sentence labels derived from a mapping against domain-specific dictionaries combined with CLEVER (weak supervision). These sentence vectors are used to train a machine learning model to determine whether UI and BD were affirmed or negated, and whether the clinician discussed risk with the patient. Finally, we combine the sentence-level annotations using majority voting to assign a unique label for the entire clinical note. For performance assessment, we compare the sentence-level classification performance against a popular generative models for text sentiment analysis: Naive Bayes model's (NB) and note-level performance against a domain-specific rule-based system.[17]

## METHODS

### Raw data source

With the approval of Institutional Review Board (IRB), the Stanford prostate cancer research database was used for analysis.[20] This contains electronic medical record (EMR) data from a tertiary care academic medical center on a cohort of 6595 prostate cancer patients with diagnosis from 2008 onwards, encompassing 528 362 unique clinical notes including progress notes, discharge summaries, telephone call notes, and oncology notes.

### Dictionaries

The two targeted treatment-related side effects following prostate cancer therapy are defined as:

- *UI: Urinary incontinence*, or the loss of the ability to control urination, is common in men who have had surgery or radiation for prostate cancer. There are different types of UI and differing degrees of severity and length of duration;
- *BD: bowel problems* following treatment for prostate cancer are common and include diarrhea, fecal incontinence, and rectal

bleeding, also with differing degrees of severity and length of duration.

A reference group of 3 clinical domain experts (2 urologists and 1 urology research nurse) gave us lists of terms relating to the presence of UI and BD by individually looking at 100 clinical notes that were retrieved from the Stanford prostate cancer research database. The lists were combined and an experienced urology nurse curated the final terms for UI (eg incontinence, leakage, post void dribbling) and BD (eg bowel incontinence, diarrhea, rectal bother). The final list (see Supplementary Table S1) not only contains terms from the 100 clinical notes but also includes additional terms important for capturing UI and BD that are based on the suggestions of the domain experts. Note that general urinary symptoms (eg nocturia, dysuria, hematuria) are not considered as affirmed UI, thus such terms are not included in the dictionary. The same UI dictionary was previously used to implement a rule-based information extraction system.[17]

## Annotations

In order to create a gold standard test set, 110 clinical notes were randomly selected from the entire corpus of notes. Two nurses and one medical student independently annotated 110 clinical notes with 120 sentences. The set of clinical notes used to create the dictionaries was isolated from the validation notes. Annotations were assigned in two levels—(1) *sentence-level*—raters went through the entire note, selected the sentences that discussed UI/BD, and assigned a label to each sentence; (2) *note-level*—raters looked at all the sentences that have been extracted on the sentence-level annotation phase, and assigned a label to the entire note. We present the sample distribution for both sentence- and note-level annotations in Supplementary Table S2.

The following labels were assigned, if applicable, for both UI and BD: (1) Affirmed: symptom present; (2) Negated: symptom negated; (3) Discussed Risk: clinician documented a discussion regarding risk of the symptom. Some example sentences retrieved from the clinical notes present in our dataset and the labels assigned by the human expert are presented in Table 1.

Inter-rater reliability at the sentence-level was estimated using Cohen's Kappa (Table 2). Moderately low agreements between the human raters reflects the subjectivity challenges associated with manual chart review. The main discrepancies occurred when the sentences contained contradictory information or unclear statements. Note that no predefined annotation protocol was available to the raters. The annotation was performed only depending on their clinical experience. Majority voting among the three raters was used to resolve the conflicting cases. These human annotations were only used to validate the automated annotation described below.

## Proposed pipeline

Our proposed pipeline consisted of three core components: (1) dictionary-based raw text analysis; (2) neural embedding of sentences; (3) discriminative modeling. The pipeline takes the free-text clinical narratives as input and categorizes each sentence according to whether the PCO was affirmed/negated or risk discussed. Figure 1 shows a diagram of the pipeline.

## Neural embedding of words

In the Stanford prostate cancer database (see Sec. Dataset), there are 164 different types of clinical narratives. In the preprocessing step, we applied standard NLP techniques to clean the text data and enhanced the semantic quality of the notes prior to neural embedding. We used a domain-independent Python parser for stop-word removal, stemming, and number to string conversion. Pointwise Mutual Information is used to extract the word-pairs to preserve the local dependencies using nltk library.[21] The bigrams with fewer than 500 occurrences were discarded to reduce the chance of instability caused by low word frequency count. The top 1000 bigram collocations were concatenated into a single word, eg 'low_dose', 'weak_stream'. In order to reduce variability in the terminology used in the narratives, we used the pre-existing CLEVER dictionary to map the terms with similar meaning that are often used in the clinical context, to a standard term list. For instance, {'mother', 'brother', 'wife',.} were mapped to FAMILY; {'no', 'absent', 'adequate to rule her out', .} mapped to NEGEX; {'suspicion', 'probable', 'possible',...} mapped to risk RISK; {'increase', 'invasive', 'diffuse',...} mapped to QUAL. The CLEVER terminology was constructed using a distributional semantics approach where a neural word embedding model was trained on large volume of clinical narratives derived from Stanford.[22] Then, after using the UMLS and the SPECIALIST Lexicon to identify a set of biomedical "seed" terms, statistical term expansion techniques were used to curate the similar terms list by identifying new clinical terms that shared the same contexts. This expanded dictionary derived empirically from heterogeneous types of clinical narratives will be more useful and comprehensive in the text standardization process compared to any single manually curated vocabulary. Bigram formulation using PMI and CLEVER root term mapping contributed to reducing sparsity in the vocabulary.

Total 528 162 preprocessed notes (excluding the test set) were used as input for a word2vec model[23] in order to produce the neural embeddings in an unsupervised manner. word2vec adopts distributional semantics to learn dense vector representations of all words in the preprocessed corpus by analyzing the context of terms. The word embeddings learned on a large text corpus are typically good at representing semantic similarity between similar words, since such words often occur in similar context in the text. For the word2vec training, we used the Gensim library[24] and the continuous bag of words model which represents each word in a vocabulary as a vector of floating-point numbers (or "word embeddings") by learning how to predict a "key word" given the neighboring words. No vectors were built for terms occurring fewer than 5 times in the corpus and the final vocabulary size was 111 272 words. We collected 50 randomly annotated sentences (for UI) to use for validation and selected the window size and vector dimension by performing grid search to optimize the best f1-score (see Figure 2).

## Training set creation from dictionary

In context of the current study, manual annotation of narrative sentences is not only laborious, but also extremely subjective as demonstrated by the inter-rater agreement scores (see Table 2). One of the major advantage of the proposed method is that no explicit ground truth sentence-level annotation is needed to train the supervised learning model. We employed the domain-specific dictionaries containing a set of affirmative expressions for UI and BD to build an artificial training set (see Dictionaries for details of dictionary creation). The UI dictionary contains 64 unique terms, indicative of UI, and BD dictionary contain 48 terms. Further the affirmative expressions are combined with NEGEX and RISK term from the CLEVER dictionary to create examples of nonaffirmed and risk

**Table 1.** Sample sentences and its corresponding annotation for UI and BD

| Urinary incontinence (UI) | | Bowel dysfunction (BD) | |
|---|---|---|---|
| Sentence | Label | Sentence | Label |
| Voiding history: two or more pads per day | Affirmed | Problems with diarrhea and rectal discomfort. | Affirmed |
| He does have some leakage late in the afternoon, which is particularly, worse, after drinking coffee or alcohol. | | We talked about eating tactics to help with loose stools including eating smaller, frequent meals instead of large meals. | |
| He has excellent urinary control and has been pad free. | Negated | He did have loose stool for 1 day on Thursday that has resolved. | Negated |
| Says that his urinary control is better, and that he no longer requires a pad in the evening. | | He has not had any hematuria or rectal bleeding since treatment. | |
| We did inform him that while surgery carries with it an approximately, 5–10%, risk of urinary incontinence | Discussed risk | Acute and long-term potential side effects from radiation therapy were discussed with the patient and his wife, including but not limited to: skin change, rectal bleeding, bowel and bladder toxicity. | Discussed risk |
| With surgery, the problem tends to be urinary leakage or incontinence; and with radiation therapy, it tends to be urinary urgency. | | Effects were discussed including low blood counts, fever, diarrhea, and fatigue. | |

**Table 2.** Agreement between raters in annotating 120 selected sentences for urinary incontinence and bowel dysfunction

| | Cohen-kappa score | |
|---|---|---|
| Annotators | Urinary incontinence | Bowel dysfunction |
| Rater 1, Rater 2 | 0.66 | 0.70 |
| Rater 1, Rater 3 | 0.72 | 0.72 |
| Rater 2, Rater 3 | 0.62 | 0.64 |

description expressions. Finally, these artificial expressions (for UI: $65 \times 3 = 195$ and for BD: $48 \times 3 = 144$) were exploited to create a 'weakly supervised' training set where each of them was labeled as whether it affirmed, denied, or discussed risk associated with UI/BD.

## Sentence vector creation

### Training set

We created the sentence-level embedding by weighting word vectors by the Tf-idf score. First, we computed the Tf-idf score for terms present in the domain dictionary, whereby Tf-idf is (i) highest when terms occur many times within a small number of training samples; (ii) lower when the term occurs fewer times in a training sample, or occurs in many samples; (iii) lowest when the term occurs in virtually all training samples (no discriminative power). The computed Tf-idf scores for the terms present in the UI and BD training dataset are shown in Figure 3. As seen from the diagram: *incontin*, *diaper*, *negex* and *risk*, scored highest for UI; and *diarrhea*, *stool*, *rectal*, *negex*, and *risk* scored highest for BD. The high score represents that these terms are clinically relevant and thus expected to have high discriminative power. Finally, the sentence vectors were created by combining the word vectors and weighting by the Tf-idf score of each term. Specifically, sentence vectors were computed with:

$$V_{\text{sen}} = \frac{1}{\|N\|} \sum\nolimits_{i=1}^{N} \text{TScore}_{w_i} \times V_{w_i},$$

where $N$ is the total number of terms present in the expression, $\text{TScore}_{w_i}$ is the Tf-idf score of word $w_i$ in, and $V_{w_i}$ refers to the word vector of word $w_i$.

### Testing set

We use a pretrained NLTK sentence tokenizer to identify the sentence boundaries for 528 362 clinical narratives, and then selected relevant sentences based on presence of terms in the domain-specific dictionaries (see Supplementary Table S1). We design a set of filtering rules for each domain to drop out irrelevant sentences—for example, ensuring *eye pad* or *nasal pad* were not misinterpreted for the *pad* associated with incontinence; or that *wound leakage* was not misinterpreted as *urinary leakage*.

Among 528 362 texts, our pipeline extracted a total of 9550 unique notes with 11 639 relevant sentences for UI and 2074 relevant notes for BD with unique sentence. For BD, we limited reports within 5 years of prostate cancer diagnosis since BD is a common symptom and we are focusing on BD as a side effect of prostate cancer treatment. In order to validate our sentence extraction pipeline outcome, we randomly selected 100 narratives from both cohorts and achieved 97% accuracy with manual validation. Finally, we generate sentence-level vector embeddings as described above.

## Discriminative model: supervised learning

Vector embeddings of the training expressions (described in the previous section) can be utilized to train parametric classifiers (eg logistic regression) as well as nonparametric classifiers [random forest, support vector machines, K-nearest neighbors (KNN)]. We chose to use multinomial logistic regression (also referred as maximum entropy modeling) with 5-fold cross validation on the training dataset. Classifier performance on the test set was reported. We refer to this classifier hereafter as the neural embedding model.

## Statistical analysis of results

A total of 117 expert annotated notes and corresponding sentences were used to validate the proposed model's outcome (see Sec. Annotations) and to compare the performance with pre-existing techniques. We adopted dual level performance analysis for both sentence- and note-level annotation.

### Sentence-level annotation

We compare the performance of our sentence-level annotation model with one of most popular generative models for text

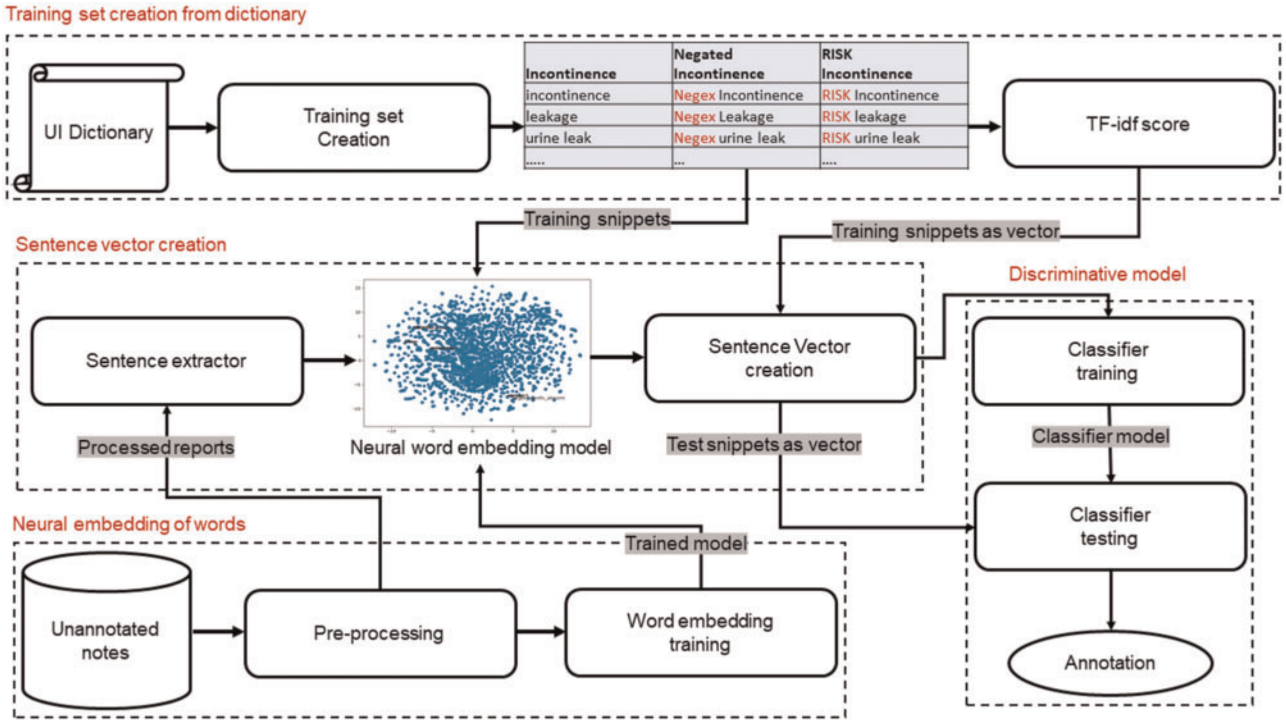
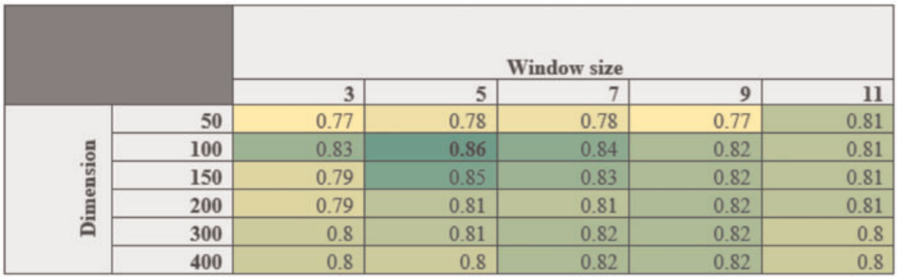


Figure 1. Pipeline for sentence-level annotation for urinary incontinence presence, absence and risk discussion. Gray highlighted texts represent I/O of the modules. Headings of the corresponding sections are mentioned along with the section numbers in red.



| | | Window size | | | | |
|---|---|---|---|---|---|---|
| | | 3 | 5 | 7 | 9 | 11 |
| Dimension | 50 | 0.77 | 0.78 | 0.78 | 0.77 | 0.81 |
| | 100 | 0.83 | **0.86** | 0.84 | 0.82 | 0.81 |
| | 150 | 0.79 | 0.85 | 0.83 | 0.82 | 0.81 |
| | 200 | 0.79 | 0.81 | 0.81 | 0.82 | 0.81 |
| | 300 | 0.8 | 0.81 | 0.82 | 0.82 | 0.8 |
| | 400 | 0.8 | 0.8 | 0.82 | 0.82 | 0.8 |

Figure 2. Validation study to optimized two hyperparameters (window size and vector dimension) for word2vec: Over all f1-score for 50 UI annotated sentences. Window size 5 and vector dimension 100 resulted best f1-score (in bold).

sentiment analysis: Naive Bayes for multinomial Bernoulli models.[13] The model estimates the conditional probability of a particular term given a class as the relative frequency of term in documents belonging to the particular class. Thus it takes into account also multiple occurrences.

### Note-level annotation

We aggregated our sentence-level annotation to the note level. Individual notes could contain multiple sentences with UI/BD related information (11 639 UI-related sentences were retrieved from 9550 notes), hence a single note may have conflicting sentence-level labels. We applied majority voting across all sentence annotations to assign a label for the note. However, we assigned priority to *Affirmed* and *Negated* labels over *Risk* labels, since clinical practitioners can discuss PCO risk in multiple sentences, but this does not confirm the current medical state of the patient.

This allowed us to compare our pipeline with the recently published rule-based method[17] for extracting UI from patient notes in prostate cancer. The rule-based method only considered affirmation and negation, so notes classified as *Discussed Risk* were grouped with the negated notes based on the absence of any positive terms.

## RESULTS

### Sentence-level annotation

Table 3 summarizes the baseline NB performance on both UI and BD test and artificial training dataset. The model achieved an average $f_1$ score of 0.57 for UI and 0.61 for BD. For the test dataset, the average precision was >0.7 but the recall remained as low as <0.55 which suggests that the comparator classifier will miss 50% information about the targeted PCOs. Table 4 summarizes the performance of our pipeline with the same training and testing datasets. Our model achieved an average $f_1$ score of 0.86 for both UI and BD, with 0.88 precision and 0.85 recall. We present the performance of both methods on the artificial train dataset to show that though the NB model was able to learn the semantics of the simple expression

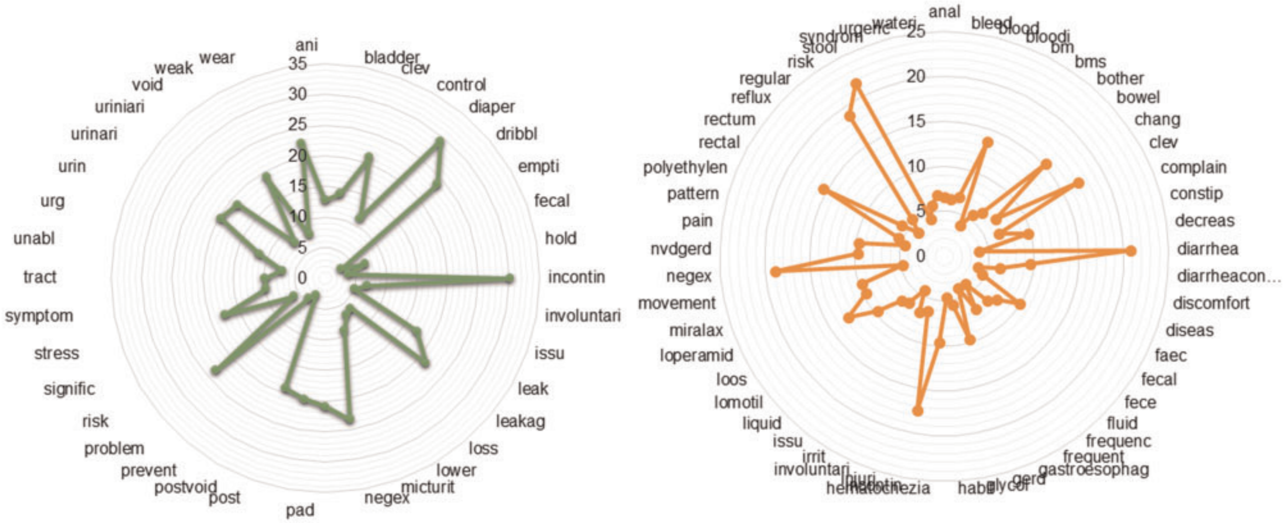


**Figure 3.** TF-IDF scores for each of the terms in the dictionaries for urinary incontinence (left) and bowel dysfunction (right).

**Table 3.** Comparator classifier's performance on the training and test datasets for UI and BD

| | Urinary incontinence | | | | | | Bowel dysfunction | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | f1-score | Precision | Recall | f1-score | Precision | Recall | f1-score | Precision | Recall | f1-score |
| | On training set | | | On test set | | | On training set | | | On test set | | |
| Affirmed | 1.00 | 1.00 | 1.00 | 1.00 | 0.44 | 0.61 | 1.00 | 1.00 | 1.00 | 0.20 | 0.50 | 0.29 |
| Negated | 1.00 | 1.00 | 1.00 | 0.25 | 0.80 | 0.38 | 1.00 | 1.00 | 1.00 | 0.90 | 0.61 | 0.73 |
| Risk | 1.00 | 1.00 | 1.00 | 0.67 | 0.55 | 0.60 | 1.00 | 1.00 | 1.00 | 0.35 | 0.47 | 0.40 |
| avg/total | 1.00 | 1.00 | **1.00** | 0.77 | 0.53 | **0.57** | 1.00 | 1.00 | **1.00** | 0.71 | 0.57 | **0.61** |

**Table 4.** Neural embedding model performance on training and test datasets for UI and BD

| | Urinary incontinence | | | | | | Bowel dysfunction | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Precision | Recall | f1-score | Precision | Recall | f1-score | Precision | Recall | f1-score | Precision | Recall | f1-score |
| | On training set | | | On test set | | | On training set | | | On test set | | |
| Affirmed | 0.93 | 0.89 | 0.91 | 1.00 | 0.81 | 0.90 | 0.88 | 0.94 | 0.91 | 0.40 | 0.67 | 0.50 |
| Negated | 0.92 | 0.88 | 0.90 | 0.50 | 0.80 | 0.62 | 0.97 | 0.94 | 0.95 | 0.85 | 0.73 | 0.79 |
| Risk | 0.92 | 0.88 | 0.90 | 0.91 | 0.91 | 0.91 | 0.97 | 0.95 | 0.96 | 0.95 | 0.91 | 0.93 |
| avg/total | 0.90 | 0.90 | **0.90** | 0.89 | 0.84 | **0.86** | 0.94 | 0.94 | **0.94** | 0.88 | 0.85 | **0.86** |

from the dictionary, it failed to interpret the complex real sentences. Whereas the proposed method being trained on the artificial training dataset, was able to classify sentences extracted from the clinical notes with morphological and syntactic word variations and show significant improvement on the test set over the NB method ($P$-value <.01).

Classification accuracy versus the Naive Bayes comparator model is shown as a confusion matrix in Figure 4. The comparator classifier tends to incorrectly predict affirmation and risk discussion in both disease states. Our neural embedding model performs significantly better in classifying negated outcomes, with an ability to classify correctly 80% UI cases and 91% BD.

We also compared our Tf-idf weighted sentence vector generation method with doc2vec paragraph embedding method (epoch 10, dimension 100, learning rate 0.02, decay 0.0002) using the same multinomial logistic regression model. However, our weighted embedding method outperformed the doc2vec since doc2vec scored 0.55 overall f1-score for UI and 0.62 overall f1-score for BD while out method scored 0.86 f1-score for both UI and BD. The modest performance of doc2vec could be due to the application of equal weight to each word rather than capturing their discriminative power in the weights.

## Note-level annotation

For the UI case-study, we consider the 117 manually annotated notes to compare performance with rule-based method where the rules was formulated with the help of Stanford Urology experts.[17] Figure 5 shows the performance of our pipeline in terms of $f_1$ score, precision, and recall compared to the rule-based model. Our model had $f_1$ score of 0.9 versus 0.49 for the rule-based model, and higher precision and recall for both Affirmed incontinence and Negated.

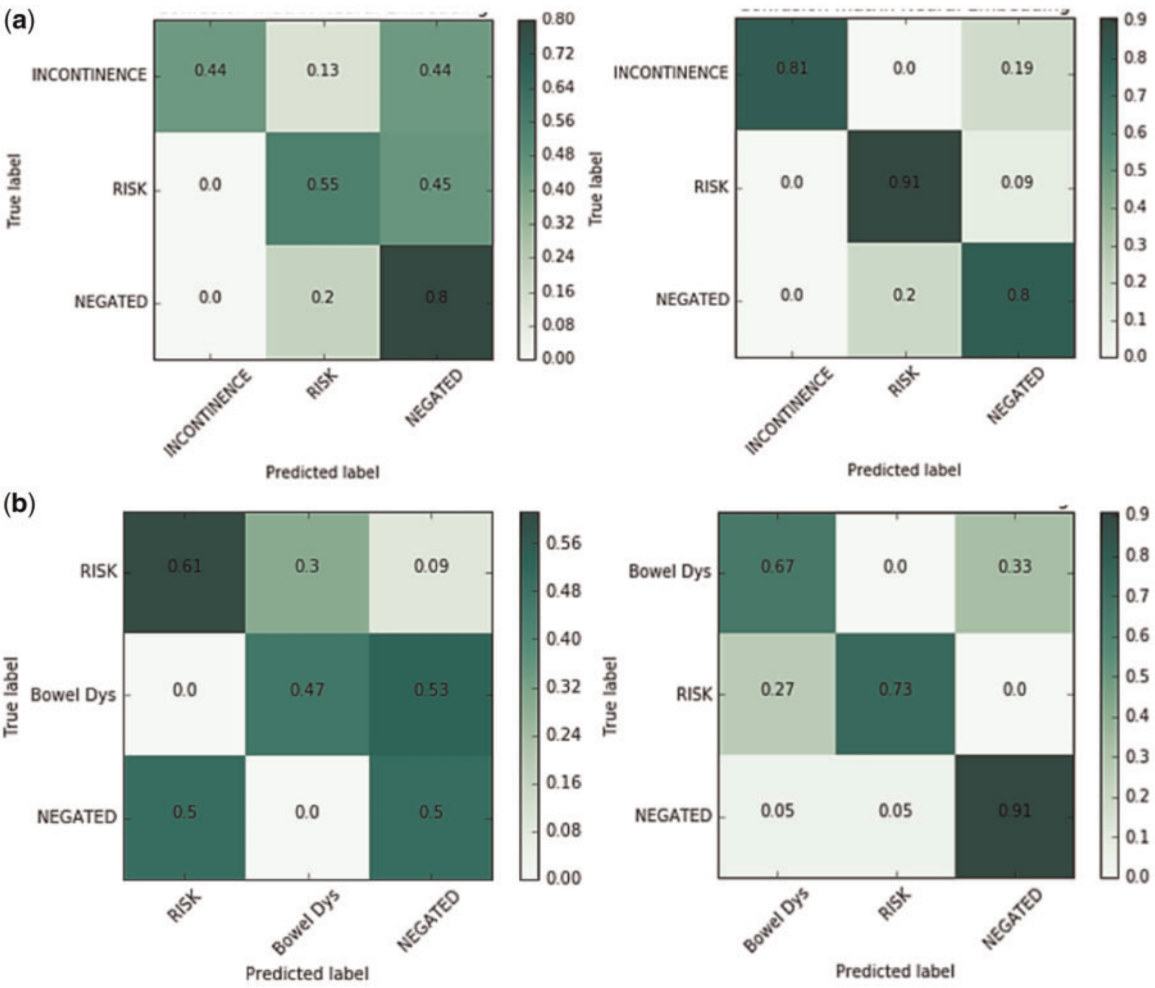


**Figure 4.** Confusion matrix for urinary incontinence (a) and bowel dysfunction (b): Baseline on right and Proposed model on the left. 44% incontinence statements have been misclassified by the baseline whereas only 19% misclassified by the proposed model. 53% bowel dysfunction statements have been misclassified by the baseline whereas only 9% misclassified by the proposed model.

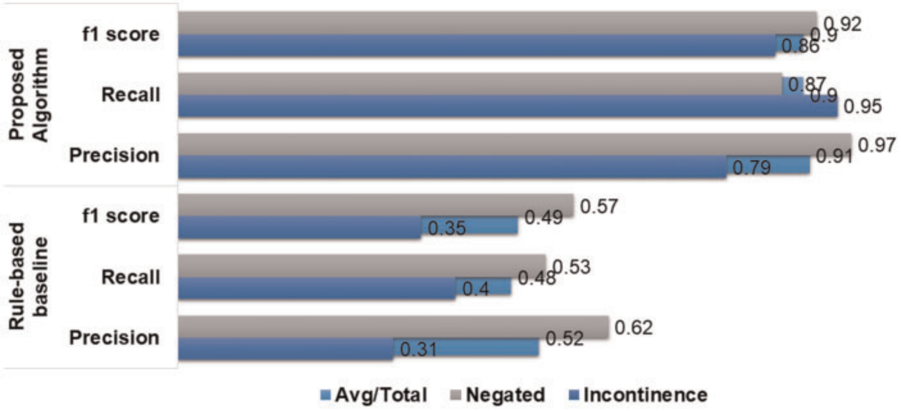


**Figure 5.** Comparative performance analysis with state-of-the-art rule-based system: urinary incontinence.

The limited performance of the rule-based method is mainly due to the concrete nature of the hard extraction rules that restricts the system to extract right information from the notes which were written using different styling/formats/expressions, even though all the notes belongs to the same institution from which the experts were involved in developing the rules. In contrast, the proposed model's performance is superior for both Affirmed and Negated incontinence, which shows the fact that the classifier trained on the

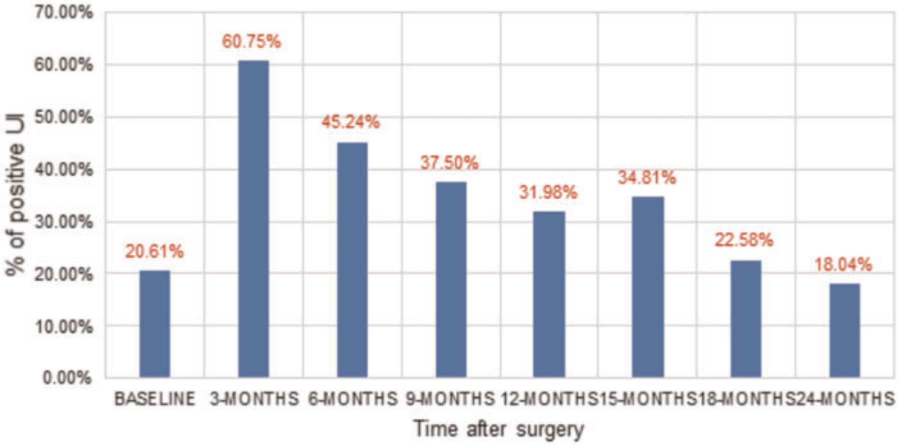


**Figure 6.** UI evaluation for radial prostatectomy patients before (BASELINE) and after surgery at different time points.

proposed embedding using an artificial training dataset is able to learn properly the linguistic variations of multiple types of clinical notes.

## DISCUSSION

### Contribution

In this study, we describe a weakly supervised NLP pipeline for assessing two important outcomes following treatment for prostate cancer, UI, and BD, from clinical notes in EMR data for a cohort of prostate cancer patients. To date, the evaluation of these outcomes has relied on labor and resource intensive methodologies, resulting in insufficient evidence regarding relative benefits and risks of the different treatment options, particularly in diverse practice settings and patient populations. As a result, efforts to establish guidelines for prostate cancer treatment based on these PCOs have been inconclusive.[22] The pipeline described here used pre-existing domain-specific dictionaries combined with publicly available CLEVER terminology as training labels, removing the need for manual chart review. This method achieved high accuracy and outperformed a previously developed rule-based system for prostate cancer treatment-associated UI.[11] Advancing the assessment of these outcomes to such scalable automated methodologies could significantly build desperately needed evidence on PCOs, and advance PCOs research in general.

### Significance

While survival is the ultimate treatment outcome, prostate cancer patients have over 99% 5-year survival rates for low-risk localized disease and therefore treatment-related side effects are a focus of informed decisions and treatment choice. However, while the risk of such complications plays a critical role in a patient's choice of treatment,[23] previous studies have suggested that urologists may underestimate or under-report the extent of these symptoms.[24] In addition, reported outcomes mainly come from high volume academic centers, which likely do not translate to other practice settings and patient populations. Recent efforts in prostate cancer care have therefore focused on the assessment and documentation of these outcomes to improve long-term quality of life following treatment[25] as well as promote patient engagement in medical care.[26] However, these outcomes are not captured in administrative or structured data, which greatly limits the generation of evidence and secondary analyze of them.[27] NLP methods present a way to automatically extract these outcomes data from clinical notes in a systematic and nonbias way,[9,24] which can significantly increase the amount of evidence available in these data and promote associated studies across disparate populations.

Existing methods for large-scale clinical note analysis rely on supervised learning[25,26] or a fixed set of linguistic rules,[26,27] which are both labor-intensive. Our weakly supervised approach is novel because it does not rely on manual annotation of sentences or notes. Instead, our approach exploits domain-specific vocabularies to craft a training set. In addition, the neural embedding allows for rich contextual information to be fed into the classifier for improved accuracy. We acknowledge that human effort is needed for the dictionary creation, but this effort is substantially less than the manual chart review effort and reusable to identify annotation for more cases. This approach outperformed a rule-based system for incontinence,[17] and showed good performance relative to a comparator classifier in both UI and BD. The application of this methodology to evaluate outcomes hidden in clinical free text may enabled the study of important treatment-related side effects and disease symptoms that cannot be captured as structure data and possibly enhance our understanding of these outcomes in populations who are not adequately represented in controlled trials and survey studies. In Figure 6, we quantified positive UI for 1665 radial prostatectomy patients applying the neural embedding model on the clinical notes that are documented before and after the surgery. The NLP extracted quantifications of the large cohort correlate well with recent clinical studies[33,34] conducted on diverse patient populations and practice setting.

### Limitations

First, assessing outcomes from clinical notes requires adequate documentation within the EHR. While significant variation in documentation rates likely exists across providers and systems, PCOs such as UI and BD in prostate cancer care are integral in evaluating the quality of care and therefore are routinely documented in the patient chart.[35] Second, the domain-specific dictionaries used in the current study were collected from a set of experts from the same clinical organization and therefore might not be generalizable to other healthcare settings. However, these outcomes and the terms used to report their assessment are fairly standardized in the community. The validation of the dictionaries in a different organization could enhance the accuracy of the pipeline, and we expect that performance could vary when multi-institutional free-text clinical notes are analyzed. Third, our model lacks sensitivity for word order which limits the ability of learning long-term and rotated scope of negex terms. However, our method is focused on sentence-level analysis thus it is not heavily impacted by long-term scope. Clinical practitioners often mention PCOs in multiple sentences of a clinical note, but the discussion of outcomes simultaneously with other unrelated topics in the same sentence was limited.

In future work, we will apply this model in other healthcare settings to test cross-institutional validity. This would require adaptation of the preprocessing step and possibly an update to the domain-specific dictionaries to capture terminology differences between sites. Additionally, the pipeline can be applied to other disease domains to test its generalizability. A new domain would require the development of a new dictionary. However, it may be possible to conduct clustering on a text corpus in order to generate the domain-specific dictionaries automatically without the need for a clinical review group.

## CONCLUSIONS

Based on weighted neural embedding of sentences, we propose a weakly supervised machine learning method to extract the reporting of treatment-related side effects following among prostate cancer patients from free-text clinical notes irrespective of the narrative style. Our experimental results demonstrated that performance of the proposed method is considerably superior to a domain-specific rule-based approach[11] on a single institutional dataset. We believe that our method is suitable to train a fully supervised NLP model where a domain dictionary has already been created and/or inter-rater agreement is very low. Our method is scalable for extracting PCOs from millions of clinical notes, which can help accelerate secondary use of EMRs. The NLP method can generate valuable evidence that could be used at point of care to guide clinical decision making and to study populations that are often not included in surveys and prospective studies.

## SUPPLEMENTARY MATERIAL

Supplementary material is available at *Journal of the American Medical Informatics Association* online.

## CONTRIBUTORS

IB developed the methodology and analyzed the results. KL, MS, and MF performed validation against the patient data. JDB, DR, and THB designed the study. TS and MS curated the database. IB, KL, MS, and THB were major contributors in writing the manuscript. All authors read and approved the final manuscript.

## FUNDING

This work was supported by the National Cancer Institute of the National Institutes of Health under Award Number R01CA183962.

*Conflict of interest statement*. None declared.

## REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2017. *CA Cancer J Clin* 2017; 67 (1): 7–30.
2. Hamdy FC, Donovan JL, Lane JA, *et al*. 10-year outcomes after monitoring, surgery, or radiotherapy for localized prostate cancer. *New Engl J Med* 2016; 375 (15): 1415–24.
3. Weiss NS, Hutter CM. Re: Comparative effectiveness of prostate cancer treatments: evaluating statistical adjustments for confounding in observational data. *J Natl Cancer Inst* 2011; 103 (16): 1277.
4. Frank L, Basch E, Selby JV; Patient-Centered Outcomes Research Institute. The PCORI perspective on patient-centered outcomes research. *JAMA* 2014; 312 (15): 1513–4.
5. Capurro D, Yetisgen M, van Eaton E, Black R, Tarczy-Hornoch P. Availability of structured and unstructured clinical data for comparative effectiveness research and quality improvement: a multisite assessment. *EGEMS (Wash DC)* 2014; 2 (1): 1079.
6. Chen J, Ou L, Hollis SJ. A systematic review of the impact of routine collection of patient reported outcome measures on patients, providers and health organisations in an oncologic setting. *BMC Health Serv Res* 2013; 13: 211.
7. Sieh W, *et al*. Treatment and mortality in men with localized prostate cancer: a population-based study in California. *Topcanj* 2013; 6: 1–9.
8. Selby JV, Beal AC, Frank L. The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. *JAMA* 2012; 307 (15): 1583–4.
9. D'Avolio LW, Litwin MS, Rogers SO, Bui AAT. Facilitating clinical outcomes assessment through the automated identification of quality measures for prostate cancer surgery. *J Am Med Inform Assoc* 2008; 15 (3): 341–8.
10. Litwin MS, Steinberg M, Malin J, Naitoh J, McGuigan KA. *Prostate Cancer Patient Outcomes and Choice of Providers: Development of an Infrastructure for Quality Assessment*. Santa Monica, CA: RAND CORP; 2000.
11. Kreimeyer K, Foster M, Pandey A, *et al*. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017; 73: 14–29.
12. Napolitano G, Marshall A, Hamilton P, Gavin AT. Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. *Artif Intell Med* 2016; 70: 77–83.
13. Skeppstedt M, Kvist M, Nilsson GH, Dalianis H. Automatic recognition of disorders, findings, pharmaceuticals and body structures from clinical text: an annotation and machine learning study. *J Biomed Inform* 2014; 49: 148–58.
14. Pons E, Braun LMM, Hunink MGM, Kors JA. Natural language processing in radiology: a systematic review. *Radiology* 2016; 279 (2): 329–43.
15. Meystre SM, *et al*. Congestive heart failure information extraction framework for automated treatment performance measures assessment. *J Am Med Inform Assoc,* 2016.
16. Hernandez-Boussard T, Tamang S, Blayney D, Brooks J, Shah N. New paradigms for patient-centered outcomes research in electronic medical records: an example of detecting urinary incontinence following prostatectomy. *EGEMS* 2016; 4 (3): 1.
17. Hernandez-Boussard T, Kourdis PD, Seto T, *et al*. Mining electronic health records to extract patient-centered outcomes following prostate cancer treatment. Presented at the AMIA Annual Symposium, 2017.
18. Gupta A, Banerjee I, Rubin DL. Automatic information extraction from unstructured mammography reports using distributed semantics. *J Biomed Inform Assoc* 2018.
19. Banerjee I, Chen MC, Lungren MP, Rubin DL. Intelligent word embeddings for radiology report annotation: benchmarking performance with state-of-the-art. *J Biomed Inform Assoc*.
20. Seneviratne M, Seto T, Blayney DW, Brooks JD, Hernandez-Boussard T. Architecture and implementation of a clinical research data warehouse for prostate cancer. *EGEMS* 2018.
21. Bouma G. Normalized (pointwise) mutual information in collocation extraction. In: Proceedings of GSCL; 2009: 31–40.

22. Tamang SR, Hernandez-Boussard T, Ross EG, Patel M, Gaskin G, Shah N. Enhanced quality measurement event detection: an application to physician reporting. *EGEMS* 2017; 5 (1): 5.
23. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Presented at the Advances in Neural Information Processing Systems 26 (NIPS 2013); 3111–3119.
24. Rehurek R, Sojka P. Software framework for topic modelling with large corpora. In: LREC 2010 Workshop on New Challenges for NLP Frameworks; 2010.
25. Wilt TJ, MacDonald R, Rutks I, Shamliyan TA, Taylor BC, Kane RL. Systematic review: comparative effectiveness and harms of treatments for clinically localized prostate cancer. *Ann Intern Med* 2008; 148 (6): 435–48.
26. Zeliadt SB, *et al*. Why do men choose one treatment over another? A review of patient decision making for localized prostate. *Cancer* 2006; 106 (9): 1865–74.
27. Litwin MS, Lubeck DP, Henning JM, Carroll PR. Differences in urologist and patient assessments of health related quality of life in men with prostate cancer: results of the CaPSURE database. *J Urol* 1998; 159 (6): 1988–92.
28. Sanda MG, Dunn RL, Michalski J, *et al*. Quality of life and satisfaction with outcome among prostate-cancer survivors. *N Engl J Med* 2008; 358 (12): 1250–61.
29. Barry MJ, Edgman-Levitan S. Shared decision making—pinnacle of patient-centered care. *N Engl J Med* 2012; 366 (9): 780–1.
30. Quan H, Li B, Duncan Saunders L, *et al*. Assessing validity of ICD-9-CM and ICD-10 administrative data in recording clinical conditions in a unique dually coded database. *Health Serv Res* 2008; 43 (4): 1424–41.
31. Murff HJ, *et al*. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA* 2011; 306 (8): 848–55.
32. Sohn S, Ye Z, Liu H, Chute CG, Kullo IJ. Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports. *AMIA Jt Summits Transl Sci Proc* 2013; 2013: 249–253.
33. Nguyen DHM, Patrick JD. Supervised machine learning and active learning in classification of radiology reports. *J Am Med Inform Assoc* 2014; 21 (5): 893–901.
34. Hripcsak G, Austin JHM, Alderson PO, Friedman C. Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology* 2002; 224 (1): 157–63.
35. Dreyer KJ, Kalra MK, Maher MM, *et al*. Application of recently developed computer algorithm for automatic classification of unstructured radiology reports: validation study. *Radiology* 2005; 234 (2): 323–9.
36. Donovan JL, Hamdy FC, Lane JA, *et al*. Patient-reported outcomes after monitoring, surgery, or radiotherapy for prostate cancer. *N Engl J Med* 2016; 375 (15): 1425–37.
37. Chen RC, Basak R, Meyer A-M, *et al*. association between choice of radical prostatectomy, external beam radiotherapy, brachytherapy, or active surveillance and patient-reported quality of life among men with localized prostate cancer. *JAMA* 2017; 317 (11): 1141–50.
38. Martin NE, Massey L, Stowell C, *et al*. Defining a standard set of patient-centered outcomes for men with localized prostate cancer. *Eur Urol* 2015; 67 (3): 460–7.