Received: 15 January 2020 | Revised: 15 June 2020 | Accepted: 23 June 2020

DOI: 10.1002/lrh2.10237

RESEARCH REPORT

Learning Health Systems

Phenotyping severity of patient-centered outcomes using clinical notes: A prostate cancer use case

Selen Bozkurt[1] | Rohan Paul[2] | Jean Coquet[1] | Ran Sun[1] | Imon Banerjee[2,3] | James D. Brooks[4] | Tina Hernandez-Boussard[1,2,5]

[1]Department of Medicine, Biomedical Informatics Research, Stanford University, Stanford, California, USA

[2]Department of Biomedical Data Sciences, Stanford University, Stanford, California, USA

[3]Department of Radiology, Stanford University, Stanford, California, USA

[4]Department of Urology, Stanford University, Stanford, California, USA

[5]Department of Surgery, Stanford University, Stanford, California, USA

Correspondence
Tina Hernandez-Boussard, Department of Medicine, Biomedical Informatics Research, Stanford University, 1265 Welch Road, #245, Stanford, CA 94305-5246, USA.
Email: boussard@stanford.edu

Funding information
National Institutes of Health, Grant/Award Number: R01CA183962; National Cancer Institute

Abstract

Introduction: A learning health system (LHS) must improve care in ways that are meaningful to patients, integrating patient-centered outcomes (PCOs) into core infrastructure. PCOs are common following cancer treatment, such as urinary incontinence (UI) following prostatectomy. However, PCOs are not systematically recorded because they can only be described by the patient, are subjective and captured as unstructured text in the electronic health record (EHR). Therefore, PCOs pose significant challenges for phenotyping patients. Here, we present a natural language processing (NLP) approach for phenotyping patients with UI to classify their disease into severity subtypes, which can increase opportunities to provide precision-based therapy and promote a value-based delivery system.

Methods: Patients undergoing prostate cancer treatment from 2008 to 2018 were identified at an academic medical center. Using a hybrid NLP pipeline that combines rule-based and deep learning methodologies, we classified positive UI cases as mild, moderate, and severe by mining clinical notes.

Results: The rule-based model accurately classified UI into disease severity categories (accuracy: 0.86), which outperformed the deep learning model (accuracy: 0.73). In the deep learning model, the recall rates for mild and moderate group were higher than the precision rate (0.78 and 0.79, respectively). A hybrid model that combined both methods did not improve the accuracy of the rule-based model but did outperform the deep learning model (accuracy: 0.75).

Conclusion: Phenotyping patients based on indication and severity of PCOs is essential to advance a patient centered LHS. EHRs contain valuable information on PCOs

Selen Bozkurt and Rohan Paul shared first authorship.

and by using NLP methods, it is feasible to accurately and efficiently phenotype PCO severity. Phenotyping must extend beyond the identification of disease to provide classification of disease severity that can be used to guide treatment and inform shared decision-making. Our methods demonstrate a path to a patient centered LHS that could advance precision medicine.

KEYWORDS

deep phenotyping, natural language processing, prostate cancer, urinary incontinence

1 | INTRODUCTION

In a learning health system (LHS), the patient is at the center of the delivery system.[1] Personalized care is enabled by harnessing data from all similar patients, which helps to understand and guide treatment decisions and value-based care. While several health systems are realizing the promise of continuous learning, the incorporation of patient-centered outcomes (PCOs) into the LHS is limited.[2-4] PCOs are difficult to capture because they can only be described by the patient, are subjective, and often documented only as unstructured text in the electronic health record (EHR).[5] Given these issues and the complexity of PCOs, computerized methods, including machine learning and natural language processing (NLP), are necessary to unlock the wealth of information buried in unstructured textual notes that often document PCOs.[6,7] In fact, using computational methods for clinical phenotyping has created opportunities to expand population-wide assessments of both PCOs and therefore patient-valued care. Maximizing the use of all data available in the EHR is fundamental to the creation of LHS and its goal of best care.[8]

Computational patient phenotyping is a data-driven task of discovering clinical phenotypes in a large patient cohort to assist clinical representations for patient groups sharing the same set of diseases.[9] The majority of data in EHRs are collected as unstructured data, as caregivers need the flexibility to freely write intuitions, possibilities, and develop narratives of the patient's disease and progression.[10] Fortunately, recent advances in technologies have provided opportunities to use information from EHR unstructured data for PCO research.[11,12] In particular, advanced machine learning and deep learning-based NLP approaches have demonstrated high accuracy to extract important PCOs to help streamline the information reported in the clinical narratives to be utilized for the LHS.[6,11,13,14] However, these methods are limited in that they only provide binary classification of symptoms (positive/negative).

PCOs include common side effects of diverse treatment regimens, which are often not investigated by population level as they are generally reported in unstructured text. Our group tackles this challenge and has focused on PCOs surrounding treatments for cancer and particularly prostate cancer. Many men treated surgically for prostate cancer experience some level of urinary incontinence (UI) during their treatment journey.[15] While some patients report minimal disruption of their quality of life due to mild UI, moderate and severe UI often have a significant negative impact on quality of life.[16] Therefore, it is important to quantify the severity of this symptom to accurately guide treatment—which can range from surgery for severe UI, to occasional use of protective pads for mild UI. Severity classification allows for triaging patients in order to identify those who might benefit from additional medical care, such as medications or surgical procedures (e.g. artificial urinary sphincter) for severe urinary symptoms.[17] Development of a severity classifier could significantly improve patient counseling by allowing both patients and providers to understand the extent and degree of side effects, such as UI, to guide treatment choice. However, without population-level studies on PCO severity, opportunities to personalize the treatment of PCOs, such as UI, are limited.

2 | QUESTION(S) OF INTEREST OR RESEARCH INTERESTS

In this study, as an example of extracting PCOs embedded in clinical text, we propose a NLP pipeline to identify the severity of UI in prostate cancer patients using only free-text clinical notes from EHRs (e.g. progress notes, nursing notes, oncology notes). We include a rule-based model, a deep learning model and a hybrid model which combines traditional rule-based NLP methods with recent advances—distributed word representations[18] and convolutional neural network (CNN) architecture[19] to accurately tackle the challenging task of classifying notes into various severities of UI. We present the performance of both rule-based and deep learning models and compare these models with the hybrid architecture. This work provides a framework to augment the LHS by ensuring patients' symptoms are part of the continuous learning from the health system—keeping the patient at the center of care.

3 | METHODS

3.1 | Data source

In this study, patients were identified in a prostate cancer clinical data warehouse (CDW), which is described in detail elsewhere.[20] In brief, data were collected from a tertiary-care academic medical center with an Epic EHR system (Epic Systems, Verona, Wisconsin) and managed in an EHR-based relational database. This study was approved by the institution's Institutional Review Board.

3.2 | Study cohort

The study included patients diagnosed with prostate cancer (ICD-9-CM:185 and ICD-10-CM:C61) between 2008 and 2018. The initial dataset consists of 1.4 million notes across 18 different note types from 15 218 patients. Patients were excluded if they did not receive initial treatment for prostate cancer (prostatectomy, radiotherapy, hormone therapy, or chemotherapy) at our institute. We further excluded patients who received cystoprostatectomy (ICD-9 procedure code: 57.71), as these patients had a primary diagnosis of bladder cancer with prostate cancer detected incidentally on surgical pathology. In addition, we excluded patients with less than two encounters throughout the study period due to limited clinical information available for these patients.

As a second cohort filtering, we used a previously described machine learning method to categorize patients into affirmed and negated classification of UI.[6] The present study used only those notes that were categorized as affirmed UI (0.90 F1 score) by this algorithm to reduce computation time. This method yielded a total of 19 213 notes of 3612 patients with predicted affirmed UI (Figure 1).

3.3 | Manually annotated reference standard data for UI severity

Since the severity of UI is inherently subjective, rules were constructed after consulting with practicing urologists and a urology nurse. These rules are summarized in Table 1.

Using the rules in Table 1, three researchers (one domain expert and two research nurses) annotated a set of sentences parsed from clinical notes based on these constructed rules. Each reviewer was given ≥200 sentences with 100 of them being common across all researchers. This enabled the calculation of interrater agreement scores while maximizing the number of labelled examples retrieved and minimizing labelling time.

Cohen's Kappa (κ) is a commonly used metric to measure agreement between two raters; a value greater than 0.8 indicates a nearly perfect agreement. The κ values for each pair of the labelers are

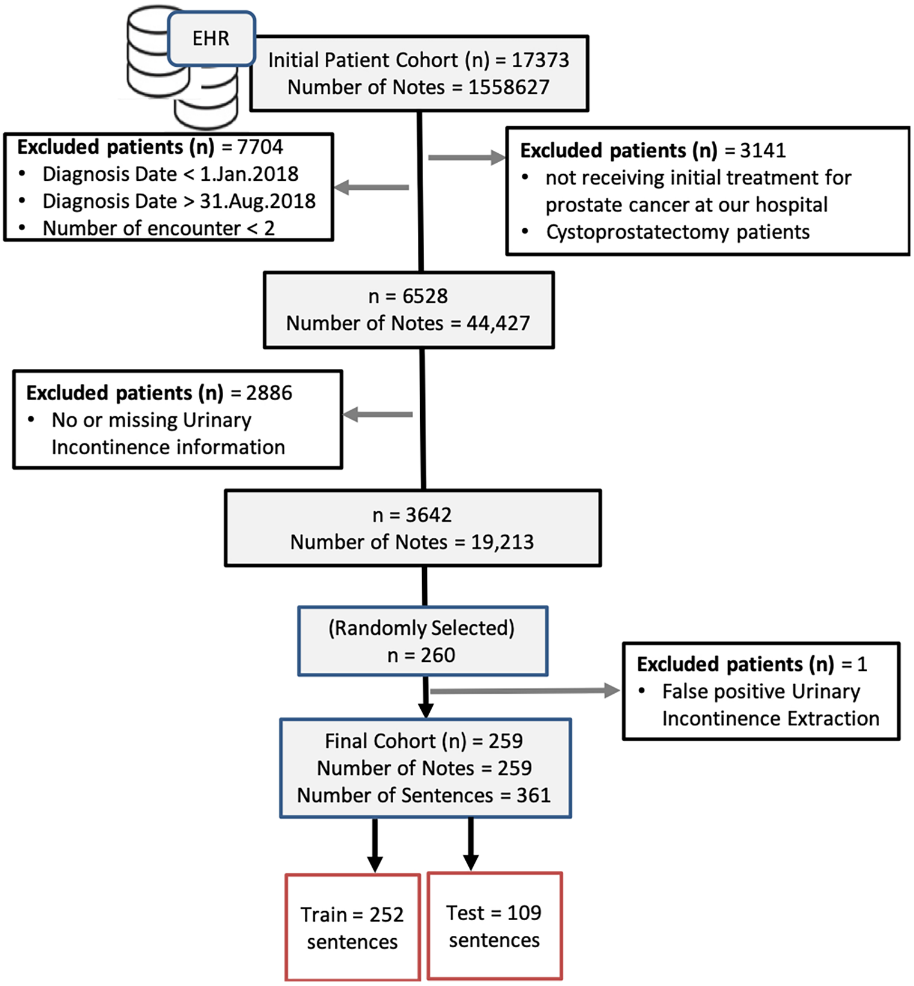

EHR
Initial Patient Cohort (n) = 17373
Number of Notes = 1558627
Excluded patients (n) = 7704
• Diagnosis Date < 1.Jan.2018
• Diagnosis Date > 31.Aug.2018
• Number of encounter < 2
Excluded patients (n) = 3141
• not receiving initial treatment for prostate cancer at our hospital
• Cystoprostatectomy patients
n = 6528
Number of Notes = 44,427
Excluded patients (n) = 2886
• No or missing Urinary Incontinence information
n = 3642
Number of Notes = 19,213
(Randomly Selected)
n = 260
Excluded patients (n) = 1
• False positive Urinary Incontinence Extraction
Final Cohort (n) = 259
Number of Notes = 259
Number of Sentences = 361
Train = 252 sentences
Test = 109 sentences

FIGURE 1 Flowchart to select the final cohort and train-test sets

TABLE 1 Rules constructed to categorize severity of urinary incontinence

	Mild	Moderate	Severe
Severity based on pad counts	≤1 pad per day	Two to three (inclusive) pads per day	>3 pads per day ≥1 diaper per day
Frequently used keywords	"Mild | minimal | occasional | rare | minor | some" used to describe incontinence Postvoid dribbling	"Moderate | considerable" used to describe incontinence	"Severe | total | complete" used to describe incontinence
Sample sentence	For example, "patient is now down to 1 pad/day"	For example, "he continues to experience moderate stress urinary incontinence"	For example, "he is totally incontinent since his surgery last month"

TABLE 2 Cohen's Kappa values to assess interrater agreement among pairs of labelers

Rater pair	κ
Rater 1, rater 2	0.867
Rater 1, rater 3	0.927
Rater 2, rater 3	0.881

shown in Table 2. The high values of κ are likely attributable to the construction of rules before the analysis stage. Labels for the common set were combined using majority voting. Labels from the remaining sentences were concatenated together to form the final dataset. This method yielded a total of 361 sentences labelled with the three severities of UI, including 132 mild, 119 severe, and 110 moderate cases.

3.4 | Hybrid deep phenotyping method—Combination of rule-based and CNN

In order to automatically classify severity of UI using clinical notes, we developed an NLP pipeline which is a combination of traditional rule-based methods and deep learning models to build a reliable classifier (Figure 2).

3.5 | Preprocessing

Notes were preprocessed using standard NLP tools that involve tokenization of sentences and words using a domain-independent parser from the NLTK package in Python and removal of the selected punctuations and extra white spaces. Through this process, a note corresponded to a list of sentences and a sentence corresponded to a list of words.

3.6 | Rule-based method

A previously developed dictionary of terms that were associated with the presence of UI was used as the knowledge base of the rule-based model.[7] The original dictionary contained 64 terms and phrases but was augmented with the addition of 6 terms for the task of detecting UI severity.

We next randomly selected a set of 20 clinical notes from the cohort to develop regular expressions used to capture the clinical concepts in the dictionary of terms. A total of seven regular expressions were formulated for mild UI, three for moderate, and six for severe. The ConText algorithm was used to detect negation, discussion of future risks, and past experiences of UI.[21] A sentence was classified as "mild" if negation was detected as "*not severe*" or none of the other rules were matched because all patient in the cohort were already classified as having some stage of UI. Sentences that matched rules for two or more severities were systematically classified as the more severe category. For example, the sentence "*He usually uses 2 pads per day*, *but recently has been using 4-5 pads*" was classified as severe even though it matches the criteria for both moderate and severe.

3.7 | CNN method

In an attempt to overcome the shortcomings of the rule-based system (e.g. limited coverage and generalizability), we developed a deep learning method to classify sentences into the three classes using an approach that utilizes a variant of max-pooling to standardize the hidden representations of sentences to the same dimensionality.[19] To address the limited number of samples for supervised training, we generated a pretrained embedding layer of the CNN model by training distributed word vectors using the word2vec algorithm for any word that appeared more than 100 times in the dataset (all 1.4 million notes).[18] To increase the generalizability of the model and to learn the complex word semantics, the entire database (prior to any filtering) was used to train the distributed word vectors. The final vocabulary included 73 024 tokens. We used the continuous bag-of-words (CBOW) variant of word2vec to train 300-dimensional word vectors. CBOW trains a supervised model to predict a central word of a fixed-sized window using the remaining words in that window. A window size of 8 was chosen since these hyperparameters have been shown to work well in other studies utilizing word vectors.[6,22] The model was trained for 10 epochs using the Gensim python library implementation of word2vec.[23] The trained vectors have the desirable property of assigning semantically similar words close together in Euclidean space. For instance, the closest token to the word "patient" was "pt,"

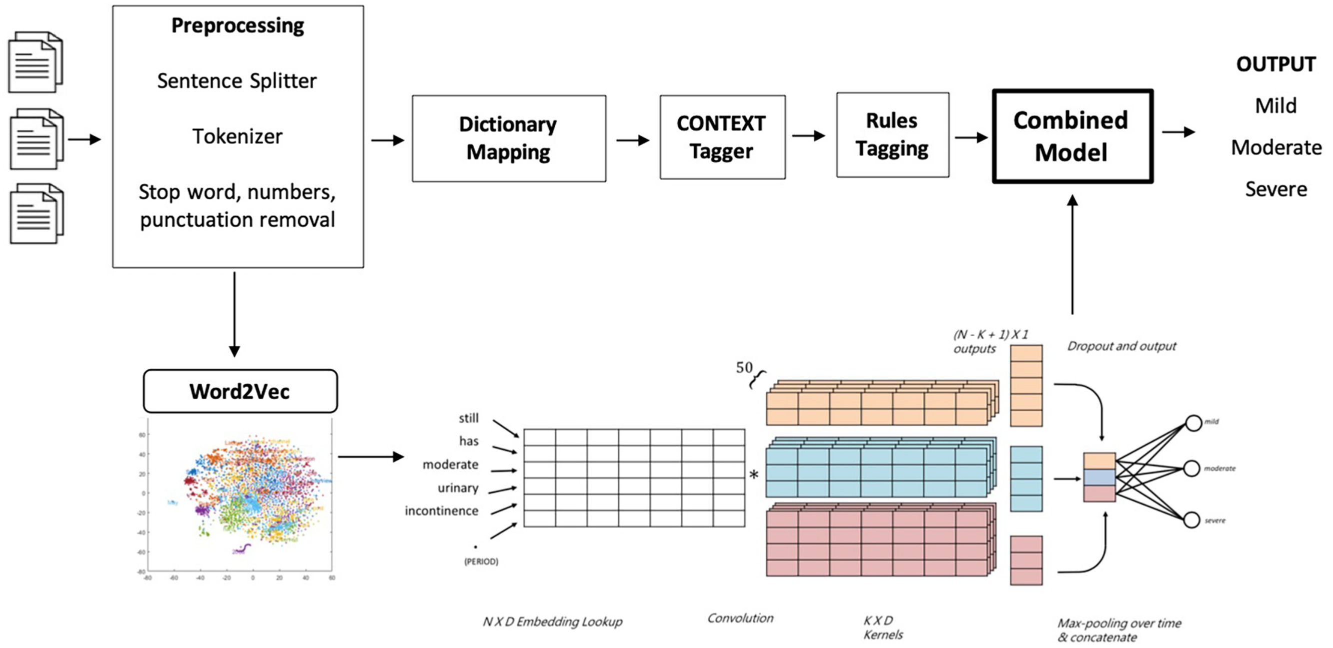


FIGURE 2 The proposed hybrid pipeline with convolutional neural network (CNN) architecture used for sentence classification. N is the number of tokens in a given sentence, D is the embedding size of 300. K is the size of a particular kernel

a commonly used clinical abbreviation, as measured by the cosine distance. The closest token to the misspelled word “incotninence” was the correctly spelled word “incontinence.”

The CNN model architecture is depicted in Figure 2. Tokenized sentences from the training set (n = 288) were the input of the model. For each word in a sentence, D = 300-dimensional pretrained word embeddings were concatenated to form the input matrix. The final model included kernels of sizes 2, 3, and 4. The outputs of these vectors were fed through a rectified linear unit nonlinearity and then “*pooled over time*” (ie, the max of each kernel output was taken to obtain a single number). These pooled vectors were concatenated to form a 50 × 3 = 150-dimensional vector. Thus, every sentence, regardless of length, was represented by a single vector with 150 entries. A dropout layer was applied with a drop probability of 0.75. Finally, the output layer consisted of a softmax nonlinearity applied to four neurons, one for each severity. The model was trained using a weighted categorical cross entropy loss function and Adam optimizer for 10 epochs using the pytorch framework.

Example sentences (n = 361), previously annotated by researchers, were split into train and test sets in a 70:30 ratio. The CNN model was trained on the training set with fivefold cross-validation. The hold out test set was used to evaluate the model performance.

3.8 | Combination of the models for deep phenotyping

Both methods were combined to obtain a single model where the target was to obtain a comprehensive representation of textual expression. Each regular expression from the rule-based model was treated as a binary feature: 1 if the pattern matched, all nonmatches scored 0. A binary vector of 19 features (16 regular expressions, negation, historical and future context discussion) was concatenated with the outcome of final hidden layer of the trained CNN model and pass through the softmax function to predict the labels as described in the CNN model.

3.9 | Evaluation

All three methods were evaluated on the same test set. Precision, recall, and F1-score were collected for each class along with an overall accuracy score. We also displayed the confusion matrix to show interclass misclassification rate.

4 | RESULTS

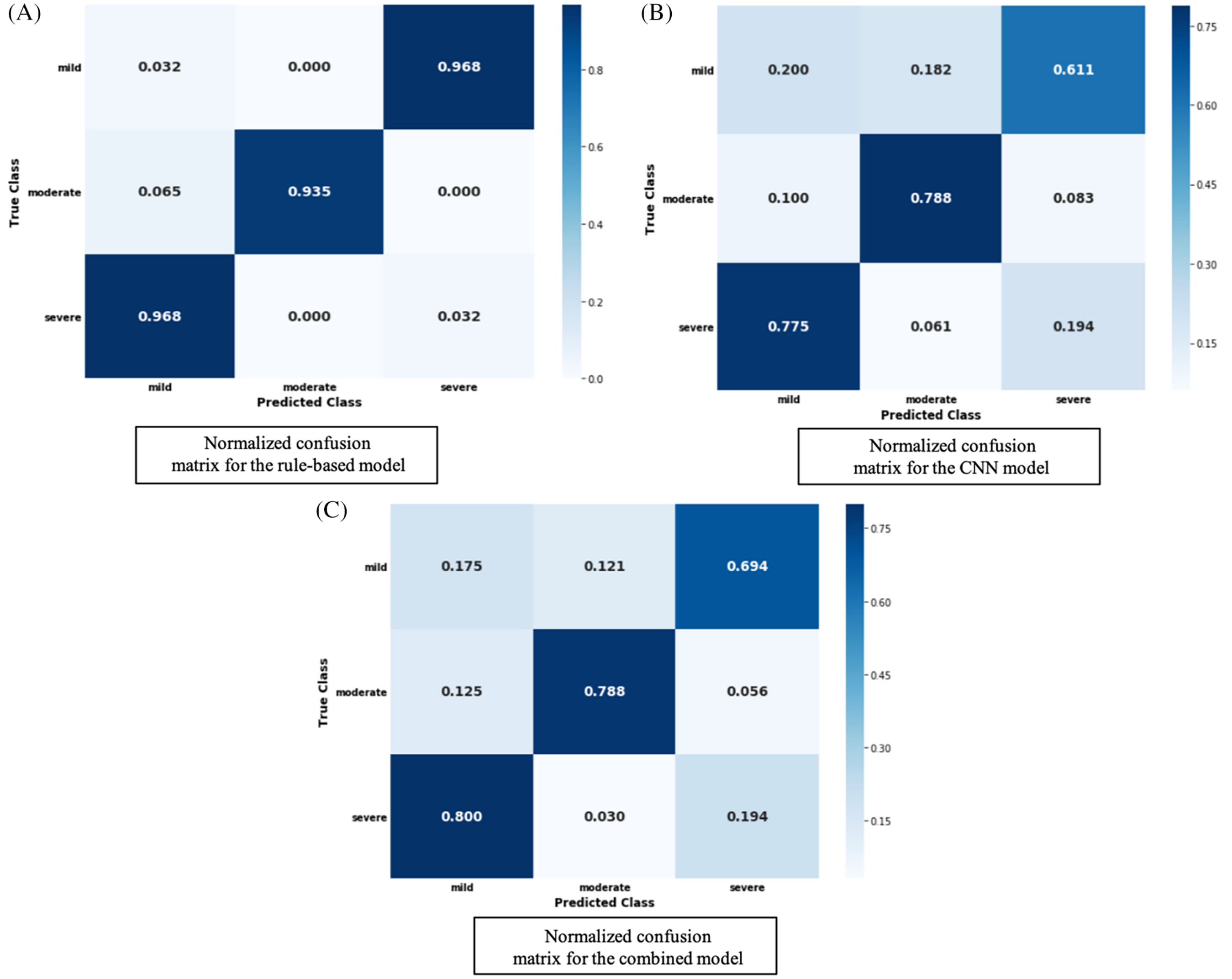
The study consisted of 259 patients with reported UI as mild (n = 87, 33.6%), moderate (n = 79, 30.5%), and severe (n = 93, 35.9%). Models were trained and tested using this cohort. The test performance metrics are summarized in Table 3, which include precision, recall, and F1-score and a confusion matrix (Figure 3). The rule-based model achieved an overall accuracy of 0.86, as compared to 0.73 in the CNN model and 0.75 in the hybrid model.

The rule-based model showed high precision but lower recall for each severity class. The model performed better in moderate and severe categories than in the mild category. The confusion matrix is presented in Figure 3A.

TABLE 3 Test set performance metrics for the rule-based and CNN models

	Rule-based model			CNN model			Hybrid model		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
Mild	0.91	0.79	0.82	0.72	0.78	0.75	0.73	0.80	0.76
Moderate	0.99	0.88	0.94	0.76	0.79	0.78	0.84	0.79	0.81
Severe	0.97	0.83	0.90	0.69	0.61	0.88	0.74	0.69	0.71

Abbreviation: CNN, convolutional neural network.



Normalized confusion matrix for the rule-based model

Normalized confusion matrix for the CNN model

Normalized confusion matrix for the combined model

FIGURE 3 Normalized confusion matrix for the different models: A, rule-based model; B, convolutional neural network (CNN) model; and C, combined hybrid model

The rule-based method had explicit rules for handling concepts specifically identified in the domain dictionary, such as the examples in Table 1. The CNN model's results underperform the rule-based model on the same test set with an overall accuracy of 0.73.

The performance metrics for the deep learning model are summarized in Table 3 and Figure 3B. Although the number of examples in the test set was small, the structure of the confusion matrix demonstrated that the two models had "learned" fundamentally distinct features for severity classification. In contrast to the rule-based model, CNN model's recall rates for mild and moderate group were higher than the precision rate. In addition, CNN model was less successful at identifying the "severe" category compared to the mild and moderate

UI categories. Combining the rule-based and CNN methods did not improve the accuracy of rule-based model (Table 3 and Figure 3C) but did slightly outperform CNN-only model with an overall accuracy of 0.75.

5 | DISCUSSION

Precision medicine is evolving from a concept under healthcare reform to a reality under an LHS framework. At the center of an LHS is evidence to guide patient-centered care, which should include predicting symptom severity, particularly for PCOs that affect quality of life, such as UI following surgical treatment for prostate cancer. While a first step is indeed to identify whether a patient has a specific symptom or not, the next step requires classifying symptoms based on their severity to guide treatment options and value-based care. We previously developed an NLP phenotyping algorithm to identify the patients with UI documented in the EHR.[6,24] Here, we build on this work by phenotyping patients based on UI severity, which classifies a positive UI into mild, moderate, and severe incontinence. Using both rule-based and deep learning NLP methods, we developed a reliable classifier for UI severity. The successful implementation of automated PCO stratification models coupled with best practices could enable a healthcare system that is patient centric and supports clinical decision-making at the point of care.

Both rule-based and machine learning NLP approaches to leverage granular data in EHRs are common and their accuracy has been demonstrated in many recent studies.[5,6,22] In our study, the rule-based approach depends on human expertise outperformed the machine learning approach on UI severity extraction task. As reported in other studies,[9,22,23,25-27] the underlying reason for this may be that the hand-designed rules that precisely capture specific patterns overfit with the data. Therefore, such rule-based approaches might not generalize well to other health care systems or other conditions because even slight variations from the rules can dramatically erode performance of the model. On the other hand, while this study does depend on these subjective rules, it provides a framework for comparing relative severities of UI within a single medical center, and, to a lesser degree, between medical centers.

The knowledge base of both methods is a domain-specific dictionary. The dictionary must be identified a priori and, unfortunately, free-text clinical documentation of UI might not be limited to the specific terms we have included since all clinicians are unlikely to use the same terms and might have local differences in terminology that could affect model performance. For example, the word *depends* can be used both as a name brand for diapers (ie, the patient wears *Depends* for UI) or as a verb (ie, the patient's UI *depends* on physical activity). To address this potential shortcoming, we have shared our dictionaries through national repositories (pheKB.org), welcome the addition of other terms and concepts, and encourage other groups to adapt the dictionaries to their specific practice settings.

Use of deep learning models for NLP tasks has shown excellent performance in mining the EHR and can contribute greatly to the analysis of unstructured text.[22,28,29] While recurrent neural network architectures are commonly used for deep learning tasks involving text classification, the convolutional approach used in our study is ideally suited for the task of parsing clinical texts as trained kernels to evaluate the independent impact of small groups of words on the target label.[19,22,29] However, our results showed that the CNN model showed poorer performance characteristics compared to the rule-based model for extraction and tabulating UI severity. The lower performance might be attributed to the paucity of training set which includes less than 100 examples for each class. Limitations of machine learning methods on small sample set has emerged as a significant challenge in recent studies, necessarily due to limited population sizes available is single EHRs where studies are interrogating single diseases, symptoms, or outcomes.[30,31] Hence, researchers have gravitated toward create hybrid frameworks using traditional NLP and machine learning solutions.[31]

To leverage the advantages of both approaches, we propose a combination of traditional rule-based methods and recent developments in deep learning to build reliable classifiers. Although the performance of rule-based system outperforms the other models, as in the case of phenotyping UI severity, it is possible that the superior performance might be due to overfitting based on selection of terms specifically used by providers in our health care system. Our hybrid NLP pipeline leveraged the complementary strengths of both the rule-based and the CNN models in classifying the severity of UI and produced higher performance than CNN only approach. Therefore, we believe our semiautomated iterative approach might produce optimal and replicable results in other settings.[31] Moreover, since the terms for UI severity classification in the model are not unique to prostate cancer patients, the model likely could be used to identify UI and its severity in men and women. Since UI is a common finding with increasing age due to pelvic floor disorders in women and benign prostatic hyperplasia in men our algorithm could have broad applications.

It has been reported that most patient data resides in unstructured clinical narratives.[32] Harvesting and organizing this valuable data stream poses significant challenges for knowledge delivery in an LHS. In addition, utilizing this information plays a significant role in data-driven and evidence-based decision-making. Fortunately, as we show in this study, recent advances in NLP methodologies tackle this challenge and streamline the acquisition of this critical information so that it can be utilized by the LHS.[8,13] For example, Kaggal et al developed an NLP infrastructure that allows data collection and large-scale analytics as a part of an LHS cycle that has been used in discrete settings to provide real-time care recommendations to clinicians at the point of care.[33] Our work, using UI as an example, shows how these discrete applications can continue to be refined. By identifying and classifying symptom or disease severity, an LHS can provide recommendations tailored to disease severity. In the case of UI, mild symptoms could elicit recommendations for behavioral-based therapies (eg, Kegel exercises or biofeedback) while moderate-severe symptoms might recommend referral for medical or surgical approaches (eg, anticholinergic medications, artificial urinary sphincter). Therefore, NLP empowered LHS infrastructures have the potential to advance health care by generating and implementing a seamless patient centered information extraction framework from entire EHR.

The LHS is a paradigm in which data generated in routine clinical care can be used for the collaborative healthcare choices of each patient and provider.[34,35] However, to understand the best treatment for any particular patient, granular details on patients' values and symptoms are needed. This entails having the knowledge from the patient to personalize care plans that meet individuals' specific needs and values rather than care plans developed for the *average patient*. However, to support a patient-centric learning healthcare system, the systematic extraction of disease severity and its association with quality of life are essential. Here, we purpose different NLP methodologies that may be used to improve the depth of information available from clinical text related to treatment-related such outcomes. To be useful to providers, the systems must incorporate relevant features of disease severity in order to personalize recommendations for management approaches. This will require direct input from the providers and likely multiple iterations to refine which information is extracted in order to provide the most useful output for the clinician. As we move to an LHS, we must be able to phenotype disease and symptom severity to move data to knowledge and knowledge to the point of care. Phenotyping the severity of treatment-acquired side effects following prostate cancer therapy is particularly compelling given that prostate cancer is the most common cancer among men in the United States and the majority of men must choose between different treatment modalities, each with important symptoms attributable to the therapy selected that affect their quality of life. Furthermore, the ability to stratify patients based on PCO severity opens opportunities for future research using biological markers, including anatomic differences observed on imaging, or DNA variations. Capturing and utilizing data embedded in clinical narratives for patient phenotyping can enhance both the rigor and the relevance of evidence in the cycle of LHS. A future challenge in implementing an LHS will be to develop methods for shortening the window of iteration that refine which PROs are captured and how they are quantified to provide more patient specific recommendations to physicians in real time.

5.1 | Limitations

Our study does have limitations. First, the NLP based phenotyping task was developed at a single academic medical center and, as mentioned above, might not be generalizable to other sites and settings due to differences in terminology. Second, our approach depends solely on sentence level information to make predictions. While information in our dataset tends to be concisely recorded, there are instances where mentions of UI can span several sentences and would be missed due to sentence-level tokenization. Therefore, the model may suffer if labels are aggregated to the note level. If, however, all sentences referring to UI mention at least one of the terms from the dictionary, they would be adequately captured by the model we have developed. Finally, our study is limited by the small number of annotated notes used for training and testing, and our models might miss some relevant but rare expressions used to identify and characterize UI in the entire EHR.

6 | CONCLUSION

A successful LHS transforms data from routine care into evidence and provides optimal care to each individual patient. In order to learn from patients' outcomes, particularly PCOs, robust patient phenotyping is needed to classify patients' disease severity. Symptom severity poses extraordinary challenges for a precision medicine approach, but its classification is essential to learn through care delivery and can ultimately improve the care provided to patients. In our study of UI, a rule-based phenotyping approach outperformed other advanced phenotyping methodologies for severity phenotyping. These results highlight the importance of understanding the phenotyping challenge and suggest that, for certain phenotypes, a rule-based approach is the best choice for discriminating disease severity. We believe that accurate phenotyping using EHRs is necessary to bring evidence to the point of care and that the best methodology for phenotyping is dependent on the specific type of evidence or content under investigation.

ACKNOWLEDGEMENTS

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.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

ORCID

Tina Hernandez-Boussard https://orcid.org/0000-0001-6553-3455

REFERENCES

1. Finney Rutten LJ, Alexander A, Embi PJ, et al. Patient-centered network of learning health systems: developing a resource for clinical translational research. *J Clin Transl Sci*. 2017;1(1):40-44.
2. Harle CA, Lipori G, Hurley RW. Collecting, integrating, and disseminating patient-reported outcomes for research in a learning healthcare system. *eGEMs*. 2016;4(1):1240.
3. Stephens KA, Osterhage KP, Fiore-Gartland B, Lovins TL, Keppel GA, Kim KK. Examining the needs of patient stakeholders as research partners in health data networks for translational research. *AMIA Jt Summits Transl Sci Proc*. 2019;2019:363-369.
4. Fagotto E Exchanging Information to Create a Learning Health System; 2019.
5. 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):1231.
6. Banerjee I, Li K, Seneviratne M, et al. Weakly supervised natural language processing for assessing patient-centered outcome following prostate cancer treatment. *JAMIA Open*. 2019;2(1):150-159.

7. Parthipan A, Banerjee I, Humphreys K, et al. Predicting inadequate postoperative pain management in depressed patients: a machine learning approach. *PLoS One*. 2019;14(2):e0210575.
8. Maddox TM, Matheny MA. Natural language processing and the promise of big data: small step forward, but many miles to go. *Circ Cardiovasc Qual Outcomes*. 2015;8(5):463-465.
9. Banda JM, Seneviratne M, Hernandez-Boussard T, Shah NH. Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu Rev Biomed Data Sci*. 2018;1:53-68.
10. Malmasi S, Ge W, Hosomura N, Turchin A. Comparison of natural language processing techniques in analysis of sparse clinical data: insulin decline by patients. *AMIA Jt Summits Transl Sci Proc*. 2019;2019:610-619.
11. Kaggal V. Learning Healthcare System Enabled by Real-Time Knowledge Extraction from Text Data. 2019.
12. Li K, Banerjee I, Magnani CJ, Blayney DW, Brooks JD, Hernandez-Boussard T. Clinical documentation to predict factors associated with urinary incontinence following prostatectomy for prostate cancer. *Res Rep Urol*. 2020;12:7-14.
13. Afzal N, Mallipeddi VP, Sohn S, et al. Natural language processing of clinical notes for identification of critical limb ischemia. *Int J Med Inform*. 2018;111:83-89.
14. Maddox TM, Albert NM, Borden WB, et al. The learning healthcare system and cardiovascular care: a scientific statement from the American Heart Association. *Circulation*. 2017;135(14):e826-e857.
15. Barocas DA, Alvarez J, Resnick MJ, et al. Association between radiation therapy, surgery, or observation for localized prostate cancer and patient-reported outcomes after 3 years. *Jama*. 2017;317(11):1126-1140.
16. Sanda MG, Dunn RL, Michalski J, et al. Quality of life and satisfaction with outcome among prostate-cancer survivors. *New Engl J Med*. 2008;358(12):1250-1261.
17. Moore KC, Lucas MG. Management of male urinary incontinence. *Indian J Urol*. 2010;26(2):236-244.
18. Turner CA, Jacobs AD, Marques CK, et al. Word2Vec inversion and traditional text classifiers for phenotyping lupus. *BMC Med Inform Decis Mak*. 2017;17(1):126.
19. Kim Y. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*. 2014.
20. Seneviratne MG, Seto T, Blayney DW, Brooks JD, Hernandez-Boussard T. Architecture and implementation of a clinical research data warehouse for prostate cancer. *EGEMS*. 2018;6(1):13.
21. Chapman WW, Chu D, Dowling JN. ConText: An algorithm for identifying contextual features from clinical text. Paper presented at: Proceedings of the workshop on BioNLP 2007: biological, translational, and clinical language processing; 2007.
22. Coquet J, Bozkurt S, Kan KM, et al. Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients. *J Biomed Inform*. 2019;94:103184.
23. Zeng Z, Deng Y, Li X, Naumann T, Luo Y. Natural language processing for EHR-based computational phenotyping. *IEEE/ACM Trans Comput Biol Bioinform*. 2019;16(1):139-153.
24. Hernandez-Boussard T, Kourdis PD, Seto T, et al. Mining electronic health records to extract patient-centered outcomes following prostate cancer treatment. *AMIA Annu Symp Proc*. 2017;2017:876-882.
25. Nguyen AN, Lawley MJ, Hansen DP, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assoc*. 2010;17(4):440-445.
26. Spasić I, Livsey J, Keane JA, Nenadić G. Text mining of cancer-related information: review of current status and future directions. *Int J Med Inform*. 2014;83(9):605-623.
27. Tan WK, Hassanpour S, Heagerty PJ, et al. Comparison of natural language processing rules-based and machine-learning systems to identify lumbar spine imaging findings related to low back pain. *Acad Radiol*. 2018;25:1422-1432.
28. Gupta A, Banerjee I, Rubin DL. Automatic information extraction from unstructured mammography reports using distributed semantics. *J Biomed Inform*. 2018;78:78-86.
29. Gehrmann S, Dernoncourt F, Li Y, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS One*. 2018;13(2):e0192360.
30. Sharma H, Mao C, Zhang Y, et al. Developing a portable natural language processing based phenotyping system. *BMC Med Inform Decis Mak*. 2019;19(Suppl 3):78.
31. Trivedi HM, Panahiazar M, Liang A, et al. Large scale semi-automated labeling of routine free-text clinical records for deep learning. *J Digit Imaging*. 2019;32(1):30-37.
32. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform 17.01*. 2008;128-144.
33. Kaggal VC, Elayavilli RK, Mehrabi S, et al. Toward a learning healthcare system—knowledge delivery at the point of care empowered by big data and NLP. *Biomed Inform Insights*. 2016;8(Suppl 1):13-22.
34. Budrionis A, Bellika JG. The learning healthcare system: where are we now? A systematic review. *J Biomed Inform*. 2016;64:87-92.
35. Friedman CP, Wong AK, Blumenthal D. Achieving a nationwide learning health system. *Sci Transl Med*. 2010;2(57):57cm29.

How to cite this article: Bozkurt S, Paul R, Coquet J, et al. Phenotyping severity of patient-centered outcomes using clinical notes: A prostate cancer use case. *Learn Health Sys*. 2020;4:e10237. https://doi.org/10.1002/lrh2.10237