DeepPhe-CR: Natural Language Processing Software Services for Cancer Registrar Case Abstraction

Harry Hochheiser, PhD, MS[1,2]; Sean Finan, BS[3,4]; Zhou Yuan, MS[1]; Eric B. Durbin, DrPH, MS[5,6]; Jong Cheol Jeong, PhD[6]; Isaac Hands, MPH[5,6]; David Rust, MS[5]; Ramakanth Kavuluru, PhD[6]; Xiao-Cheng Wu, MD, MPH[7]; Jeremy L. Warner, MD, MS, FAMIA, FASCO[8,9]; and Guergana Savova, PhD, MS[3,4]

DOI https://doi.org/10.1200/CCI.23.00156

ABSTRACT

PURPOSE Manual extraction of case details from patient records for cancer surveillance is a resource-intensive task. Natural Language Processing (NLP) techniques have been proposed for automating the identification of key details in clinical notes. Our goal was to develop NLP application programming interfaces (APIs) for integration into cancer registry data abstraction tools in a computer-assisted abstraction setting.

METHODS We used cancer registry manual abstraction processes to guide the design of DeepPhe-CR, a web-based NLP service API. The coding of key variables was performed through NLP methods validated using established workflows. A container-based implementation of the NLP methods and the supporting infrastructure was developed. Existing registry data abstraction software was modified to include results from DeepPhe-CR. An initial usability study with data registrars provided early validation of the feasibility of the DeepPhe-CR tools.

RESULTS API calls support submission of single documents and summarization of cases across one or more documents. The container-based implementation uses a REST router to handle requests and support a graph database for storing results. NLP modules extract topography, histology, behavior, laterality, and grade at 0.79-1.00 F1 across multiple cancer types (breast, prostate, lung, colorectal, ovary, and pediatric brain) from data of two population-based cancer registries. Usability study participants were able to use the tool effectively and expressed interest in the tool.

CONCLUSION The DeepPhe-CR system provides an architecture for building cancer-specific NLP tools directly into registrar workflows in a computer-assisted abstraction setting. Improved user interactions in client tools may be needed to realize the potential of these approaches.

ACCOMPANYING CONTENT

Data Supplement

Accepted October 4, 2023
Published December 19, 2023

JCO Clin Cancer Inform 7:e2300156

INTRODUCTION

Effective cancer surveillance presents a challenge in coordinated data collection. Cancer abstracts often begin at a hospital where a patient is diagnosed with or treated for cancer. This initial hospital abstract is collected by skilled registrars, often holding the Certified Tumor Registrar (CTR) certification, according to constantly updated guidelines and data standards published by organizations such as the National Cancer Institute's SEER program, the Center for Disease Control's National Program of Cancer Registries (NPCR), and the American College of Surgeons' Commission on Cancer (CoC). Cancer registrars are expected to synthesize information from a variety of clinical notes, pathology reports, and other sources of narrative text for submission to many different local and national agencies, according to increasingly complex data collection standards.

Partially or fully automated approaches based on natural language processing (NLP) techniques have been implemented as a means of increasing efficiency and accuracy (as reviewed by Savova, et al[1] and Wang et al[2]). A variety of NLP methods have been applied to a number of challenges in extracting cancer information from clinical text, including treatments,[3] recurrences,[4] and other attributes.[5] Recent efforts have explored the application of these techniques to SEER registry tasks, including the application of multitask convolutional neural networks (CNNs) to documents from the Louisiana Tumor Registry,[6] the use of CNNs to extract information from pediatric cancer pathology notes,[7] mapping of concepts in pathology notes to International Classification of Diseases (ICD)-O-3 codes,[8] and the use of transfer learning techniques to apply models trained in one registry to notes from other registries.[9]

CONTEXT

Key Objective

To develop software services for integration of Natural Language Processing (NLP) techniques for extracting cancer-relevant data into cancer registry data abstraction tools.

Knowledge Generated

NLP tools effectively (F1 >0.75) extract cancer attributes from cancer clinical notes. A containerized implementation provides application programming interface access suitable for integrating NLP functionality into cancer registry tools. A preliminary user study of a state-based registry tool augmented with these tools suggests that these approaches might be useful to registry staff.

Relevance

DeepPhe-CR is a natural language processing tool that uses rule-based machine learning and ontologies to automate extraction of data for state cancer registry reporting across multiple tumor types. Usability testing suggests a high degree of accuracy although improvement in registrar work efficiency will require further refinement and experience.

Several tools have used NLP techniques to extract phenotypes from the EHR clinical narrative.[1] Recent efforts by Alawad et al[6] and Yoon et al[10] have successfully applied multitask deep convolutional learning to extract attributes including site, laterality, behavior, histology, and grade. However, these tools work in batch modes, processing large corpora of textual documents and assembling outputs for a subsequent analysis. These methods also have relied on a bronze standard for ground truth where the predictions are modeled using the abstracted values and the collection of pathology documents for a tumor, rather than gold-standard codes derived from the pathology documents (one or many) available at the time of abstraction. This approach is not without limitations: performance is good for more common cancer sites but suffers for rarer sites.

Since 2014, we have been developing the Deep Phenotyping for Oncology Research (DeepPhe) system to extract and visualize longitudinal patient histories from clinical text to combine with the structured data from electronic medical records (EMRs). Based on the Apache Clinical Text and Knowledge Extraction System (cTAKES) NLP platform,[11] DeepPhe combines NLP for identifying and coding key cancer variables with capabilities for summarization to provide views at multiple granularities, from individual mentions to high-level summaries.[12] We have adapted our DeepPhe system to build Deep Phenotyping for Cancer Registries (DeepPhe-CR),[13] a software-service platform for embedding cancer-optimized NLP into cancer registrar data abstraction tools and workflows in a computer-assisted setting. DeepPhe-CR provides REST[14] Application-Programming Interface (API) calls designed to support integration with registry software in a computer-assisted setting. Here, we describe the goals and design of the system and lessons learned through the interactions with registrars at multiple SEER registries.

METHODS

Basic Requirements

As the widely used SEER*DMS software and other prominent registry tools are web-based, we used the REST software architectural style[14] to develop a client-side application programming interface (API) to enable registry tools to communicate with the NLP services. We exposed DeepPhe-CR functionality via API calls for submitting documents and retrieving results, including the type and location of relevant text spans.

Primary site (topography, major and minor; NAACCR Item No. 400), histology (NAACCR Item No. 522), behavior (NAACCR Item No. 523), laterality (NAACCR Item No. 410), grade (NAACCR Item No. 440), and stage (NAACCR Item Nos. 1001, 1002, 1003, 1004 for clinical TNM and 1011, 1012, 1013, 1014 for Pathologic TNM) were identified as the key variables to be extracted from the notes. DeepPhe-CR also extracts certain biomarkers. In discussions with the SEER program at the NCI, we established an initial goal of extraction of each of these attributes at F1 scores of 0.75 or better to demonstrate initial feasibility for computer-assisted abstraction. This goal of 0.75 F1 is guidance provided by the Cancer Registries on the basis of their previous work on efficiency within a computer-assisted setting; full automation requires 0.95 F1. F1 is the harmonic mean of precision/positive predictive value and recall/sensitivity and is the classic metric for reporting overall NLP system performance.

Information Extraction

An NLP system supporting the single-document registry workflow was used to develop the approaches for extracting the required data items. The DeepPhe-CR NLP system was built on the Apache cTAKES platform with extensive augmentation. Standard cTAKES analytic components on the basis of both rules and machine learning algorithms were used in the resolution of sentences, tokens, parts of speech, and named entity attributes.[11] DeepPhe-CR rule-based components and a custom cancer ontology incorporating ICD-O were used for section and table identification, named entity recognition, entity attribute assignment, entity normalization, entity relation extraction, concept summarization, and cancer attribute resolution and normalization. The DeepPhe-CR algorithms include similarity-based lookup and expressions for named entity recognition guided by an ontology, a set of rules on the basis of syntactic function and context for entity attribute assignment, and rules combined with directed graph traversal for concept summarization. Therefore, the methods across the modules range from rule-based (eg, negation) to machine learning (eg, sentence boundary tagger, part-of-speech tagger, negation) and knowledge-based (similarity-based lookup guided by an ontology). Some of the modules combine rule-based and machine learning methods (eg, negation). Thus, DeepPhe-CR implements a pipeline approach for processing the text.

Two domain experts and an informatician developed detailed annotation guidelines and piloted them on a set of 30 patients. Disagreements were tracked and discussed, and the annotation guidelines were adjusted to address areas of disagreement. Interannotator agreement was computed at kappa = 0.77-1 on an additional set of 30 patients. Gold-standard annotations for 1,560 randomly selected patients with several common cancer types—breast, lung, colorectal, ovarian, and prostate as well as a rare type of cancer (pediatric brain cancer)—were created on the basis of the finalized annotation guidelines and split into training, development, and test sets. The data were obtained from two SEER cancer registries (Kentucky and Louisiana). Table 1 provides the distribution across types of cancers and train/development/test splits.

Integration Example, User Study, and Generalization

The DeepPhe-CR tools were integrated into the Kentucky Cancer Registry's Cancer Patient Data Management System (CPDMS), a production software platform developed and implemented across the state.[15] This initial implementation was designed to demonstrate the feasibility of including DeepPhe-CR in a registry-scale software system in a computer-assisted abstraction setting. In the usability study, we used the implementation for document-level abstraction (although the system can perform summarization over multiple documents).

In a series of recorded sessions, three members of the project team who were skilled with cancer data abstraction were provided a brief introduction to the augmented CPDMS tool. They then used both the original CPDMS tool and the enhanced version to abstract the information for cancer cases, each supported by documents of varying lengths and relevant context. Accuracy, task completion time, and qualitative observations were used to descriptively assess the utility of the revised workflow.

Discussions with registry staff from two additional states (Massachusetts and Louisiana) were used to verify the generalizability of the workflow to new contexts.

Ethics Statement

This study was approved by the Human Research Protection Office of the University of Pittsburgh (STUDY19020173), the Boston Children's Hospital Institutional Review Board (M10-06-0269), and the University of Kentucky Institutional Review Board (IRB 50835).

Data Availability

The data set used in this study is not publicly available because of institutional restrictions associated with privacy risks of sharing clinical notes. DeepPhe-CR source code is available at GitHub.[13]

RESULTS

Inquiries, API Development, and Prototype

Our inquiries into the registry workflows revealed that registry processes are driven by individual documents, not patients. Thus, it is desirable for the DeepPhe-CR API to include calls for submitting individual documents and receiving appropriate results. We developed a Docker-based architecture involving two containers: a reverse proxy for managing requests, which are then handed off to the document-processing container. This container includes several key components found in the document-processing

TABLE 1. Data Distribution (No. of patients) Across Common Types of Cancers (breast, lung, prostate, colorectal, ovarian) and a Rare Type of Cancer (pediatric brain cancer)

Patient Set	Breast	Lung	Prostate	Colorectal	Ovarian	Pediatric Brain	Total
Train	231	198	178	90	90	120	907
Development	72	55	48	30	29	41	234
Test	134	120	106	30	29	39	419
Total	437	373	332	150	148	200	1,560

container: (1) core NLP functionality, (2) a summarizer module, and (3) a query processor capable of handling requests for summarization and other stored data. An overview of the DeepPhe-CR architecture is given in the Data Supplement (Fig S1).

The DeepPhe-CR API provides four calls. The most basic call summarizes a single document. Two calls allow submission of a document to be included in a future summary for a patient: one version immediately returns the results for that document, while retaining information extracted from that document for an eventual summary. The alternative version simply returns an acknowledgment without any document-level information. The final call supports retrieval of all documents submitted for a given individual. Sample calls and returned results are given in the Data Supplement (Appendix B).

DeepPhe-CR's Docker-based containerized implementation is straightforward to install. DeepPhe-CR source code (all open source), installation instructions, documentation API calls, and tools for managing the Docker containers, including execution of integration tests, can be found on the DeepPhe-CR release GitHub site.[13]

Information Extraction

DeepPhe-CR results on the held-out test split are provided in Table 2. DeepPhe-CR achieves high F1 scores on both the common (colorectal, prostate, breast, ovary, and lung) and rare (pediatric brain) types of cancers. Because DeepPhe-CR assigns a value for each required category of topography, histology, behavior, laterality, and grade, the F1 equates to accuracy, and thus, precision and recall values are the same (assuming no missing values in the gold annotations). DeepPhe-CR achieved 0.90 F1 for biomarker extraction with precision/positive predictive value 0.91 and recall/sensitivity 0.89 on the test split from Table 1.

Integration Example, User Study, and Generalization

Two features were added to the CPDMS abstraction screen to integrate DeepPhe-CR: (1) controls displaying suggested items extracted by DeepPhe-CR and (2) a text box displaying the document from which information is extracted, with spans highlighted in colors matching those used to label the suggested items. Selected items can be copied to the data annotation inputs with a single click (Fig 1).

In an initial usability study, two registrars experienced with cancer data abstraction used the enhanced CPDMS/DeepPhe-CR tool to annotate cancer documents. Both were able to use the tool appropriately and expressed enthusiasm for the enhancements.

For one of the two participants, task completion time was significantly faster with the DeepPhe-CR annotations (with DeepPhe-CR, average time 31.7 seconds, standard deviation [SD] 14.0 seconds; without DeepPhe-CR, average time 59.7 seconds, SD 14.9 seconds, Wilcoxon's W, *P* < .05). No significant difference was found for the other participant.

DISCUSSION

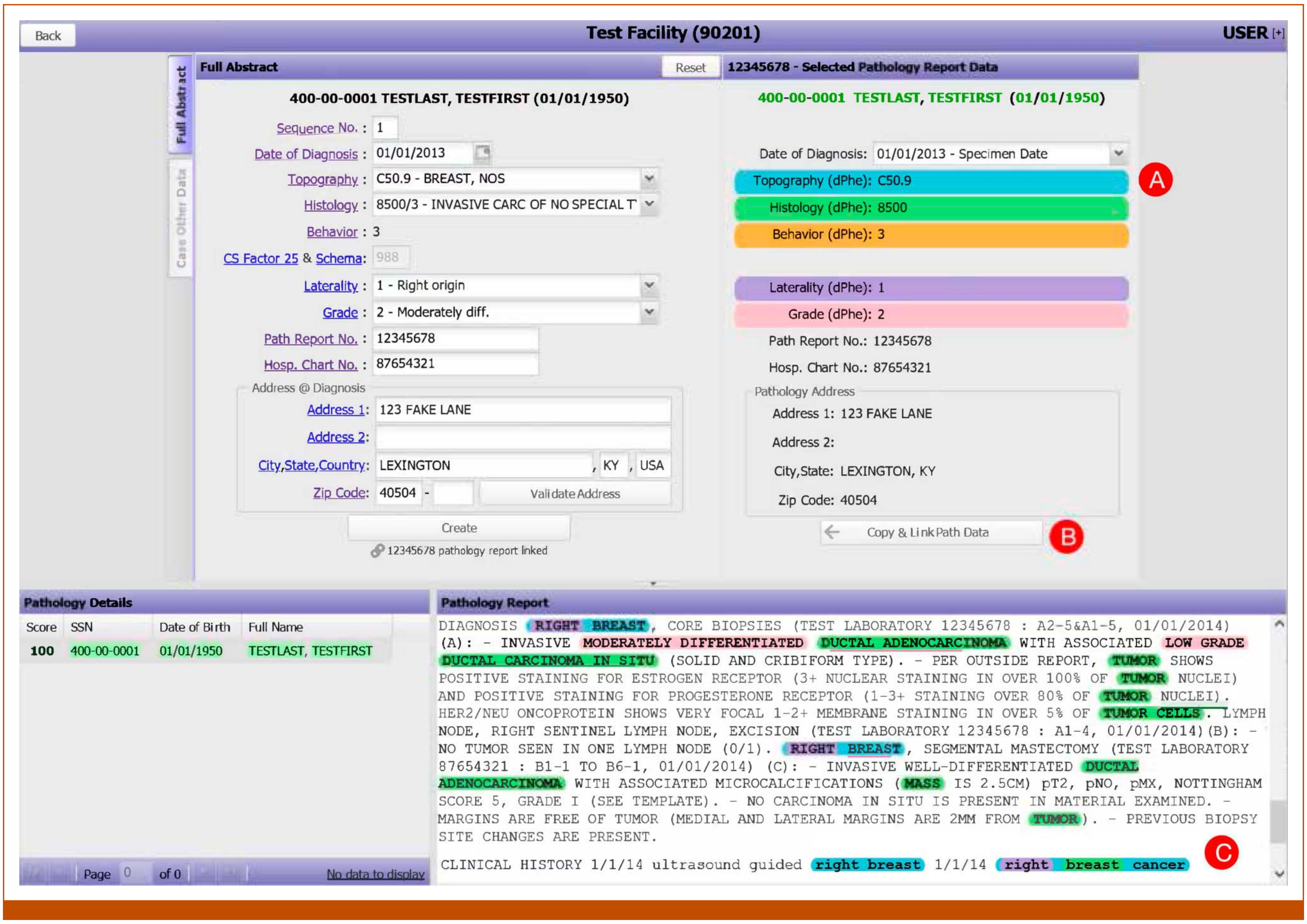
Cancer surveillance is built on the careful interpretation of clinical free text, a resource-intensive process, requiring significant person power from trained experts. NLP tools capable of automating or semiautomating the identification of these key details can be used to improve registrar efficiency and accuracy. Improving the efficiency of registry workflows also facilitates the expansion of the registry data set to include additional information, such as genomic biomarkers.

Our DeepPhe-CR system provides an architecture for building cancer-specific NLP tools directly into registrar workflows in a computer-assisted abstraction setting. Based

TABLE 2. DeepPhe-CR Results for Cancer Core Attributes on the Held-Out Test Split

	Colorectal, Prostate, Breast, Ovary, and Lung Cancers			Pediatric Brain Cancer		
Attribute	*P*	*R*	F1	*P*	R	F1
Topography: ICD-O code (major site)	.91	0.91	0.91	1.00	1.00	1.00
Topography: ICD-O code (subsite)	.85	0.85	0.85	.79	0.79	0.79
Histology: ICD-O code	.83	0.83	0.83	.87	0.87	0.87
Behavior: ICD-O code	.88	0.88	0.88	.97	0.97	0.97
Laterality: ICD-O code	.96	0.96	0.96	.97	0.97	0.97
Grade: ICD-O code	.81	0.81	0.81	.92	0.92	0.92
AJCC Pathologic T value	.98	0.92	0.94	NA	NA	NA
AJCC Pathologic N value	.98	0.98	0.98	NA	NA	NA
AJCC Pathologic M value	.99	1.00	0.99	NA	NA	NA

Abbreviations: AJCC, American Joint Committee on Cancer; F1, harmonic Mean of Precision and Recall; ICD, International Classification of Diseases; M, metastasis; N, node; P, precision/positive predictive value; R, recall/sensitivity; T, tumor.



Test Facility (90201)

Full Abstract

400-00-0001 TESTLAST, TESTFIRST (01/01/1950)

Sequence No. : 1
Date of Diagnosis : 01/01/2013
Topography : C50.9 - BREAST, NOS
Histology : 8500/3 - INVASIVE CARC OF NO SPECIAL T
Behavior : 3
CS Factor 25 & Schema: 988
Laterality : 1 - Right origin
Grade : 2 - Moderately diff.
Path Report No. : 12345678
Hosp. Chart No. : 87654321

Address @ Diagnosis
Address 1: 123 FAKE LANE
Address 2:
City,State,Country: LEXINGTON , KY , USA
Zip Code: 40504 -

12345678 pathology report linked

12345678 - Selected Pathology Report Data

400-00-0001 TESTLAST, TESTFIRST (01/01/1950)

Date of Diagnosis: 01/01/2013 - Specimen Date
Topography (dPhe): C50.9
Histology (dPhe): 8500
Behavior (dPhe): 3
Laterality (dPhe): 1
Grade (dPhe): 2
Path Report No.: 12345678
Hosp. Chart No.: 87654321

Pathology Address
Address 1: 123 FAKE LANE
Address 2:
City,State: LEXINGTON, KY
Zip Code: 40504

Pathology Details

Score	SSN	Date of Birth	Full Name
100	400-00-0001	01/01/1950	TESTLAST, TESTFIRST

Pathology Report

```
DIAGNOSIS RIGHT BREAST, CORE BIOPSIES (TEST LABORATORY 12345678 : A2-5&A1-5, 01/01/2014)
(A): - INVASIVE MODERATELY DIFFERENTIATED DUCTAL ADENOCARCINOMA WITH ASSOCIATED LOW GRADE
DUCTAL CARCINOMA IN SITU (SOLID AND CRIBIFORM TYPE). - PER OUTSIDE REPORT, TUMOR SHOWS
POSITIVE STAINING FOR ESTROGEN RECEPTOR (3+ NUCLEAR STAINING IN OVER 100% OF TUMOR NUCLEI)
AND POSITIVE STAINING FOR PROGESTERONE RECEPTOR (1-3+ STAINING OVER 80% OF TUMOR NUCLEI).
HER2/NEU ONCOPROTEIN SHOWS VERY FOCAL 1-2+ MEMBRANE STAINING IN OVER 5% OF TUMOR CELLS. LYMPH
NODE, RIGHT SENTINEL LYMPH NODE, EXCISION (TEST LABORATORY 12345678 : A1-4, 01/01/2014)(B): -
NO TUMOR SEEN IN ONE LYMPH NODE (0/1). RIGHT BREAST, SEGMENTAL MASTECTOMY (TEST LABORATORY
87654321 : B1-1 TO B6-1, 01/01/2014) (C): - INVASIVE WELL-DIFFERENTIATED DUCTAL
ADENOCARCINOMA WITH ASSOCIATED MICROCALCIFICATIONS (MASS IS 2.5CM) pT2, pNO, pMX, NOTTINGHAM
SCORE 5, GRADE I (SEE TEMPLATE). - NO CARCINOMA IN SITU IS PRESENT IN MATERIAL EXAMINED. -
MARGINS ARE FREE OF TUMOR (MEDIAL AND LATERAL MARGINS ARE 2MM FROM TUMOR). - PREVIOUS BIOPSY
SITE CHANGES ARE PRESENT.

CLINICAL HISTORY 1/1/14 ultrasound guided right breast 1/1/14 right breast cancer
```

FIG 1. The original data abstraction screen from the Cancer Patient Data Management System, augmented to support suggested data items as extracted from clinical text by DeepPhe-CR: (A) document-level topography, histology, behavior, laterality, and grade values are shown color-coded and labeled with appropriate ICD-O/NAACCR codes; (B) suggested items and demographics can be copied to the in-progress abstract with a single click; (C) the clinical text is highlighted with color codes, indicating text spans associated with the five summary values displayed in the suggestion area above (A). Thus, "Right breast cancer" is associated with topography C50.9, "breast" with histology 8500, and "right" with laterality 1. ICD, International Classification of Diseases; NAACCR, North American Association of Central Cancer Registries.

on a containerized implementation and REST-like[14] architecture, these components can easily be installed and accessed by web-based registry tools, thus minimizing changes to familiar workflows and encouraging adoption. An initial deployment within the Kentucky Cancer Registry's CPDMS provides a demonstration of the feasibility of this approach.

A future goal is to develop methods to enable the near-complete automation of many cancer registry data abstraction tasks, which requires performance in the 95%-97% F1 range. Methods for identifying the high confidence high F1 extractions could be developed as functionalities within DeepPhe-CR to enable such a process. DeepPhe-CR tools provide input that helps registrars complete their work within the computer-assisted abstraction setting. We also note that DeepPhe-CR is a modular architecture with inherent pluggable utility to registry software and can be enhanced with alternative NLP methods.

The use of DeepPhe-CR annotations as a tool for supporting manual abstraction and coding is consistent with our observations during our user study of the integration of DeepPhe-CR into the Kentucky CPDMS. Although participants were enthusiastic about the NLP assistance and showed clear signs of using the extracted attributes, we did not see consistent indication that the tool helped reduce the time required to complete abstraction tasks. This appeared to be due to the need to verify system feedback as participants regularly scrolled through documents to find highlighted text corresponding to DeepPhe-CR suggestions. System suggestions that were either omitted (ie, when DeepPhe-CR was unable to identify a tumor attribute) or incorrect appeared to increase task completion difficulty as users had to read the note to find appropriate text. We believe that continuous and prolonged use of DeepPhe-CR and further methods and user interface refinements will increase familiarity of and trust in the system.

Enhancements to the registry software user interface and workflow might provide additional gains through appropriate revisions to registry workflows. Specifically, DeepPhe-CR's ability to summarize multiple pathology reports might reduce the effort needed to manually link details across reports. However, revisions to data abstraction workflows and tools will be necessary to realize these improvements.

Further work might be needed to determine appropriate approaches for evaluating the impact of an NLP-augmented abstraction process. Although reductions in the time required for document abstraction might appear to be the most obvious potential benefit, other metrics such as subjective satisfaction and abstracting accuracy should be considered. Improvements in these alternative measures might be sufficient evidence to justify adoption of the NLP-aided approach to cancer data abstraction.

Limitations of this work include the small sample size in terms of the number of participants in the usability study and the focus of the usability study on single-document abstraction. Although the document corpus includes a diverse set of cancers across two cancer registries, thus allowing for methods' generalizability, it remains possible that results on unseen data sets will not reach the F1 scores reported here. The validation of the NLP approaches with clinical data from additional registries along with additional types of cancers and of the tools with users from those sites will be needed to demonstrate even broader generalizable utility.

In conclusion, the use of NLP to assist in the extraction of required reporting of cancer attributes and streamlining cancer data abstraction processes is an appealing possibility. The DeepPhe-CR system provides a containerized set of abstraction tools supported by a REST API, providing infrastructure suitable for integration into registry software. Accurate methods and a use case demonstrate the feasibility of this approach, while also pointing to some limitations. The DeepPhe-CR tools are available at GitHub.[13] We welcome feedback from potential users.

AFFILIATIONS

[1]Department of Biomedical Informatics, University of Pittsburgh, Pittsburgh, PA
[2]Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA
[3]Boston Childrens' Hospital, Boston, MA
[4]Harvard Medical School, Boston, MA
[5]Kentucky Cancer Registry, Markey Cancer Center, Lexington, KY
[6]Division of Biomedical Informatics, College of Medicine, University of Kentucky, Lexington, KY
[7]Louisiana Cancer Registry, New Orleans, LA
[8]Lifespan Health System, Providence, RI
[9]Legorreta Cancer Center at Brown University, Providence, RI

CORRESPONDING AUTHOR

Harry Hochheiser, Department of Biomedical Informatics, University of Pittsburgh, 5607 Baum Blvd, Pittsburgh, PA 15217; Twitter: @hshoch; e-mail: harryh@pitt.edu.

PREPRINT VERSION

A preliminary version of this paper was published on medRxiv: https://doi.org/10.1101/2023.05.05.23289524.

SUPPORT

Supported by NCI grants UH3CA243120 (H.H., S.F., Z.Y., E.B.D., J.C.J., I.H., D.R., R.K., J.L.W., and G.S) and U24CA248010 (H.H., J.L.W., and G.S.) This work was also supported the Markey Cancer Center's Cancer Research Informatics Shared Resource Facility (P30 CA177558), NCI-SEER (HHSN261201800013I/HHSN26100001) and CDC-NPCR (NU58DP007144) (E.B.D., J.C.J., I.H., D.R., and R.K.). Additional support was provided by NCI-SEER (HHSN261201800007I/HHSN26100002) and CDC-NPCR (NU58DP006332) (X.-C.W.).

DATA SHARING STATEMENT

Clinical notes used in this study cannot be shared, as they may contain personally identifiable information. All source codes are available at https://deepphe.github.io.

AUTHOR CONTRIBUTIONS

Conception and design: Harry Hochheiser, Zhou Yuan, Eric B. Durbin, Jong Cheol Jeong, Isaac Hands, Jeremy L. Warner, Guergana Savova
Financial support: Harry Hochheiser, Jeremy L. Warner, Guergana Savova
Administrative support: Harry Hochheiser, Guergana Savova
Provision of study materials or patients: Harry Hochheiser, Eric B. Durbin, Jong Cheol Jeong, David Rust, Xiao-Cheng Wu
Collection and assembly of data: Harry Hochheiser, Eric B. Durbin, Jong Cheol Jeong, Isaac Hands, David Rust, Xiao-Cheng Wu, Guergana Savova
Data analysis and interpretation: Harry Hochheiser, Sean Finan, Zhou Yuan, Eric B. Durbin, Jong Cheol Jeong, Isaac Hands, Ramakanth Kavuluru, Guergana Savova
Manuscript writing: All authors
Final approval of manuscript: All authors
Accountable for all aspects of the work: All authors

AUTHORS' DISCLOSURES OF POTENTIAL CONFLICTS OF INTEREST

The following represents disclosure information provided by authors of this manuscript. All relationships are considered compensated unless otherwise noted. Relationships are self-held unless noted. I = Immediate Family Member, Inst = My Institution. Relationships may not relate to the subject matter of this manuscript. For more information about ASCO's conflict of interest policy, please refer to www.asco.org/rwc or ascopubs.org/cci/author-center.
Open Payments is a public database containing information reported by companies about payments made to US-licensed physicians (Open Payments).

Harry Hochheiser
Research Funding: Philips Respironics (Inst)

Eric B. Durbin
Travel, Accommodations, Expenses: Caris Life Sciences, Inc

Isaac Hands
Consulting or Advisory Role: Perthera
Travel, Accommodations, Expenses: Perthera

Ramakanth Kavuluru
Stock and Other Ownership Interests: Clover Health, Teladoc

Jeremy L. Warner
This author is the Editor-in-Chief of *JCO Clinical Cancer Informatics*. Journal policy recused the author from having any role in the peer review of this manuscript.
Stock and Other Ownership Interests: HemOnc.org
Consulting or Advisory Role: Westat, Flatiron Health, Melax Tech
Other Relationship: HemOnc.org

No other potential conflicts of interest were reported.

ACKNOWLEDGMENT

David Harris and Lisa Witt provided vital annotation help and assistance in user-testing prototype tools. Brent Mumphrey, Susan Gershman, and Hung Tran provided useful feedback to the development of the tools described in this article.

REFERENCES

1. Savova GK, Danciu I, Alamudun F, et al: Use of natural language processing to extract clinical cancer phenotypes from electronic medical records. Cancer Res 79:5463-5470, 2019
2. Wang L, Fu S, Wen A, et al: Assessment of electronic health record for cancer research and patient care through a scoping review of cancer natural language processing. JCO Clin Cancer Inform 10.1200/CCI.22.00006
3. Zeng J, Banerjee I, Henry AS, et al: Natural language processing to identify cancer treatments with electronic medical records. JCO Clin Cancer Inform 10.1200/CCI.20.00173
4. Karimi YH, Blayney DW, Kurian AW, et al: Development and use of natural language processing for identification of distant cancer recurrence and sites of distant recurrence using unstructured electronic health record data. JCO Clin Cancer Inform 10.1200/CCI.20.00165
5. Bitterman D, Miller T, Harris D, et al: Extracting relations between radiotherapy treatment details, in Proceedings of the 3rd Clinical Natural Language Processing Workshop. Online: Association for Computational Linguistics, 2020. pp 194-200. https://aclanthology.org/2020.clinicalnlp-1.21
6. Alawad M, Gao S, Qiu JX, et al: Automatic extraction of cancer registry reportable information from free-text pathology reports using multitask convolutional neural networks. J Am Med Inform Assoc 27:89-98, 2019
7. Yoon HJ, Peluso A, Durbin EB, et al: Automatic information extraction from childhood cancer pathology reports. JAMIA Open 5:ooac049, 2022
8. Rios A, Durbin EB, Hands I, et al: Assigning ICD-O-3 codes to pathology reports using neural multi-task training with hierarchical regularization, in Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics. New York, NY, Association for Computing Machinery, 2021. pp 1-10
9. Alawad M, Gao S, Qiu J, et al: Deep transfer learning across cancer registries for information extraction from pathology reports. 2019 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI), 2019. pp 1-4.
10. Yoon HJ, Klasky HB, Gounley JP, et al: Accelerated training of bootstrap aggregation-based deep information extraction systems from cancer pathology reports. J Biomed Inform 110:103564, 2020
11. Savova GK, Masanz JJ, Ogren PV, et al: Mayo clinical text analysis and knowledge extraction system (cTAKES): Architecture, component evaluation and applications. J Am Med Inform Assoc 17: 507-513, 2010
12. Savova GK, Tseytlin E, Finan S, et al: DeepPhe: A Natural Language Processing system for extracting cancer phenotypes from clinical records. Cancer Res 77:e115-e118, 2017
13. DeepPhe: A Natural Language Processing System for Extracting Cancer Phenotypes from Clinical Records, 2023. https://deepphe.github.io
14. Fielding RT, Taylor RN, Erenkrantz JR, et al: Reflections on the REST architectural style and "principled design of the modern web architecture" (impact paper award), in Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering. New York, NY, Association for Computing Machinery, 2017. pp. 4-14
15. Kentucky Cancer Registry: CPDMS.net Hospital Cancer Registry Management System, 2023. https://www.kcr.uky.edu/software/cpdmsnet.php