

Application Notes

Natural language processing for abstraction of cancer treatment toxicities: accuracy versus human experts

Julian C. Hong ^{1,2,3} Andrew T. Fairchild,³ Jarred P. Tanksley,³ Manisha Palta³ and Jessica D. Tenenbaum⁴

¹Department of Radiation Oncology, University of California, San Francisco, San Francisco, California, USA, ²Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, California, USA, ³Department of Radiation Oncology, Duke University, Durham, North Carolina, USA and ⁴Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, USA

Corresponding Author: Julian C. Hong, MD, MS, Department of Radiation Oncology, Bakar Computational Health Sciences Institute, University of California, San Francisco, 1825 Fourth Street, Suite L1101 San Francisco, CA 94158, USA (julian.hong@ucsf.edu)

Received 31 August 2020; Revised 26 October 2020; Editorial Decision 29 October 2020; Accepted 30 October 2020

ABSTRACT

Objectives: Expert abstraction of acute toxicities is critical in oncology research but is labor-intensive and variable. We assessed the accuracy of a natural language processing (NLP) pipeline to extract symptoms from clinical notes compared to physicians.

Materials and Methods: Two independent reviewers identified present and negated National Cancer Institute Common Terminology Criteria for Adverse Events (CTCAE) v5.0 symptoms from 100 randomly selected notes for on-treatment visits during radiation therapy with adjudication by a third reviewer. A NLP pipeline based on Apache clinical Text Analysis Knowledge Extraction System was developed and used to extract CTCAE terms. Accuracy was assessed by precision, recall, and F1.

Results: The NLP pipeline demonstrated high accuracy for common physician-abstracted symptoms, such as radiation dermatitis (F1 0.88), fatigue (0.85), and nausea (0.88). NLP had poor sensitivity for negated symptoms.

Conclusion: NLP accurately detects a subset of documented present CTCAE symptoms, though is limited for negated symptoms. It may facilitate strategies to more consistently identify toxicities during cancer therapy.

Key words: natural language processing; radiation therapy; chemoradiation; toxicity; cancer

LAY SUMMARY

Expert abstraction of acute toxicities is critical in oncology research but can be labor-intensive and highly variable. We developed and assessed a natural language processing (NLP) pipeline to extract symptoms from clinical notes in comparison to physician reviewers. NLP accurately identified documented present Common Terminology Criteria for Adverse Event symptoms but had limited detection for documented negated symptoms. Given limitations in human review, it may facilitate research strategies to more consistently identify toxicities during cancer therapy.

INTRODUCTION

The abstraction of treatment and disease-related symptomology is critical in oncology research. As prospective toxicity documentation on clinical trials underestimates adverse events, the most rigorous method integrates retrospective human review, forming the anchor of both prospective and retrospective studies in oncology.¹ However, manual review, whether by a clinician or clinical research assistant, is labor-intensive and prone to human variation.^{2,3}

This critical clinical and analytical need presents an important potential use for the implementation of natural language processing (NLP). NLP can leverage increasing computational power and electronic health records (EHRs) to automate the systematic extraction of data from free text. Clinical NLP has been an area of active interest given the expansive and important data locked exclusively in clinical free-text notes.⁴ A number of broad clinical NLP tools have been developed and are available to extract content from clinical notes, including Apache clinical Text Analysis Knowledge Extraction System (cTAKES), MetaMap, and Clinical Language Annotation, Modeling, and Processing (CLAMP) Toolkit.⁵⁻⁷ Continued evolutions in machine learning such as deep learning have subsequently facilitated specific use-cases where underlying patterns in text can be associated directly with specific concepts, such as clinical outcomes.⁸

In oncology, NLP efforts have largely focused on the extraction of data and insights from semistructured text, such as radiology⁹ and pathology¹⁰ reports. There have been limited efforts evaluating the accuracy of the extraction of toxicity data. In particular, NLP tools are limited by their gold standard corpora, and the annotations generated by a few reviewers. Adaptation and validation for specific use can enable its use in clinical research. Its implementation offers opportunities for more consistent extraction of clinical data and may also facilitate automated extraction of data to augment clinical prediction and decision support tools.^{11,12}

Given the limitations and variability in human expert review,² the objective of this study was to develop and evaluate an NLP pipeline against human expert reviewers for the extraction of the National Cancer Institute (NCI) Common Terminology Criteria for Adverse Events (CTCAE) symptoms.

METHODS

This study was approved by the Duke University Medical Center Institutional Review Board (Pro00082776). We developed an NLP pipeline based on publicly available tools for extracting CTCAE v5.0 terms from oncology notes. As previously described,² 100 randomly selected notes for weekly scheduled radiotherapy on-treatment visits (OTV) at a single academic center between 2005 and 2016 were independently reviewed by two senior radiation oncology residents.

Patients undergoing radiotherapy are seen by their physicians during weekly OTVs to manage symptoms related to treatment or disease. The documentation for these visits can be institution-specific, but are typically brief in a SOAP format, with a subjective section describing patient symptoms, an objective section including focused physical exam findings, and an assessment and plan. OTV documentation is typically captured in a medical center-wide EHR (as is the case at our institution) or in a department-centric oncology information system. At our institution, notes are primarily free-text, though standardized EHR templates prepopulate vital signs and physical exam headers, and physical exam findings can be selected

from predefined options. Style and content varied across physicians and disease sites. As with other radiation oncology notes, specialty abbreviations can be included, such as “fx,” for “fraction” (the delivery of one radiation treatment), “Gy” the abbreviation for “Gray” (a unit of radiation dose). However, language would be anticipated to be recognizable across oncologic specialties, particularly in describing symptoms. OTV notes also have a very limited automated text population in comparison to consultation or follow-up notes. Notes reviewed in this study did not include explicit structured CTCAE toxicities.

Reviewers were instructed to comprehensively identify explicitly present, negated, or not mentioned CTCAE symptoms and were blinded to each other’s labels. This was performed utilizing a checklist of all CTCAE terms, sorted by the system, available from the NCI in multiple formats.¹³

A thesaurus (previously published and embedded in available code on GitHub) was created to harmonize overlapping CTCAE terms identified by the reviewers (e.g. cough and productive cough).^{2,14} Labels were then reviewed by an attending radiation oncologist to create a consensus.

The plain text notes were processed through the open-source Apache cTAKES v4.0.0 default clinical pipeline.⁵ cTAKES consists of multiple components to process clinical free text, including a sentence boundary detector, tokenizer, normalizer, part-of-speech tagger, shallow parser, and a named entity recognition annotator with negation.⁵ The default clinical pipeline is an easily accessible deployment which includes annotations for the most commonly desired outputs.¹⁵ Among the annotations provided are anatomical sites, signs/symptoms, procedures, disease/disorders, and medications. These were initially mapped as SNOMED CT terminology and mapped to Medical Dictionary for Regulatory Activities (MedDRA) terms using the Observational Health Data Sciences and Informatics Athena vocabulary. Since v4.0, CTCAE has been integrated into MedDRA, with mapping available from the NCI.¹³ Our code for processing the cTakes extracted terms is available on GitHub.¹⁴ Given additional MedDRA terms identified by cTAKES, we generated and made publicly available a separate thesaurus to map alternative terms to corresponding CTCAE elements ([Supplementary Data](#) and available on GitHub).

NLP output was compared against human consensus. For both human and NLP abstraction, symptoms with multiple appearances in a note were designated as present if there was at least one positive mention. Standard evaluation statistics were generated, including precision (positive predictive value), recall (sensitivity), and F1 (harmonic mean of precision and recall) for individual symptoms.¹⁶ The unweighted Cohen’s kappa coefficient between NLP and each of the reviewers was also assessed to provide a broad assessment.^{17,18}

RESULTS

As previously described, 100 notes written by 15 physicians were evaluated, representing diverse disease sites ([Table 1](#)).² No notes were from the same patient or treatment course. Among the most commonly present terms on human review, such as radiation dermatitis, fatigue, nausea, pruritis, and noninfectious cystitis, NLP demonstrated overall good precision, recall, and F1 ([Table 1](#)). Of note, the NLP pipeline did not detect urinary urgency (MedDRA code 10046593). It was, however, very sensitive (1.00) for noninfectious cystitis (F1 0.75). NLP demonstrated good performance in identifying some symptoms that had previously demonstrated low human inter-rater reliability, including radiation dermatitis, fatigue, nonin-

Table 1. Note characteristics and extracted symptoms

Word count	Median 203	IQR 164.5–237.5			
Character count	Median 1324.5	IQR 1103.25–1592.5			
Number of note authors	15				
Disease site	Number (N = 100)				
Breast	32				
Head and neck	15				
Prostate	13				
Central nervous system	10				
Lung	8				
Gynecologic	7				
Bladder	4				
Metastases (spine, spine, adrenal, leg/lung)	4				
Sarcoma	3				
Esophagus	1				
Skin	1				
Pelvic lymphoma	1				
Multiple myeloma	1				
Most common present symptoms	Number present (N = 100)	Precision (PPV)	Recall (sensitivity)	F1	Reviewer Kappa
Dermatitis-radiation	35	0.97	0.80	0.88	0.57
Fatigue	34	1.00	0.74	0.85	0.51
Pain	24	0.36	0.63	0.45	0.65
Nausea	13	0.92	0.85	0.88	0.86
Pruritus	11	0.91	0.91	0.91	0.67
Cystitis, noninfectious	9	0.60	1.00	0.75	0.00
Diarrhea	8	0.28	0.63	0.38	0.92
Mucositis	8	0.83	0.63	0.71	0.62
Urinary urgency	8	NA	0.00	NA	0.83
Folliculitis	7	1.00	0.14	0.25	0.00
Hot flashes	7	0.54	1.00	0.70	0.92
Total	277				
Most common negated symptoms	Number negated (N = 100)	Precision (PPV)	Recall (sensitivity)	F1	Reviewer Kappa
Dermatitis-radiation	42	0.89	0.19	0.31	0.57
Pain	27	0.5	0.07	0.13	0.65
Superficial soft tissue fibrosis	19	NA	0.00	NA	0
Diarrhea	18	1	0.11	0.20	0.92
Seroma	18	NA	0.00	NA	0.93
Thrush	16	1	0.31	0.48	0.11
Hematuria	16	NA	0.00	NA	0.88
Hematochezia	16	1	0.06	0.12	0.93
Dysuria	15	NA	0.00	NA	0.81
Pruritis	13	1	0.85	0.92	0.67
Urinary incontinence	13	NA	0.00	NA	0.96
Total	358				

IQR: interquartile range; PPV: positive predictive value.

Number present or negated based on consensus adjudication of identifications by both reviewers, rather than the total number of times symptoms were identified by either reviewer.

fectious cystitis, and folliculitis. Precision was also more limited for documentation of pain (0.36; F1 0.45). Example NLP errors for pain and diarrhea, two more common symptoms, are presented in Table 2.

NLP was more limited in detecting negated symptoms, the most common of which were radiation dermatitis, pain, and soft tissue fibrosis (Table 1). In general for negated symptoms, NLP demon-

strated low recall, though accompanied with high precision. NLP did demonstrate strong detection for the negation of pruritis, which was noted as a negated symptom in 13 notes.

For comparison with inter-rater variability of expert abstraction, the unweighted Cohen’s kappa coefficients compared to each reviewer were 0.52 (95% confidence interval 0.49–0.56) and 0.49

Table 2. Examples of challenging note phrases for common symptoms

Note phrase
“significant pain on the right side of his face”
“instructed on soft foods and pain control for maintaining PO intake”
“she is not having any residual pain”
“she had one episode of diarrhea today”
“she has been having 5–6 loose bowel movements daily, taking 3 Imodium/day”
“diarrhea none”

(0.52–0.55). This was lower than the unweighted kappa between the two reviewers (0.68, 0.65–0.71).²

DISCUSSION

NLP offers a potential method for detecting specific documented present CTCAE v5.0 symptoms in comparison to human review. Given the effort and inter-rater variability intrinsic to the expert review, it may be a good option for systematically assessing toxicities from unstructured clinical data.² Additionally, it may also support or validate toxicities identified during clinical trials or retrospective analysis. Notably, there was greater variability between NLP and each individual reviewer than across the two reviewers. In particular, it was limited in its ability to identify expert-identified negated symptoms, a more semantically complex task. However, most research use-cases prioritize the identification of present symptoms.

NLP did demonstrate worse performance with certain symptoms, among these, notably pain (0.36 precision and 0.63 sensitivity). This may be attributable to the multitude of pain-related terms, which may reduce recognition accuracy. There were a number of false positives. These included more complex concepts such as anticipatory guidance for future symptoms—“instructed on soft foods and pain control for maintaining PO intake.”—as well as examples of missed negations—“she is not having any residual pain.” Several examples of missed identification explicitly included the word “pain,” with some demonstrating more separation between the term identifying pain and the site—for example, phrases such as “significant pain on the right side of his face.” NLP did not identify this example as general or site-specific pain concepts.

Diarrhea was another term that challenged NLP (precision 0.28 and recall 0.63). We identified simple misses—“she had one episode of diarrhea today”—as well as more ambiguous phrases—“she has been having 5–6 loose bowel movements daily, taking 3 Imodium/day.” False positives were primarily missed negations, including those for incomplete sentences like “Diarrhea none.”

In our study, NLP was compared against annotation by multiple senior radiation oncology residents, who we expect to have comparable accuracy to attending physicians given their responsibility for the majority of clinical documentation as well as their active academic engagement in evaluating studies that incorporate CTCAE. This specialty multiexpert review is also important in evaluating this specific use case, as NLP efforts, clinical trials, and retrospective studies alike frequently utilize individual clinical research assistants, medical students, or nonsubject matter experts.

NLP has had an increasing number of applications in the oncology space. There have been limited data validating the effectiveness of NLP to extract accurately named entities in comparison to human review. A number of efforts have occurred in the semistructured

space, working to extract data from pathology reports, including staging and histology,¹⁰ and in radiology, including Breast Imaging Reporting and Data System (BI-RADS) assessments. Efforts in plain text have also been utilized to identify patients with advanced and metastatic cancer.¹⁹ Importantly, the Cancer Deep Phenotype Extraction (DeepPhe) system is a cancer-centric NLP system built on cTAKES for the abstraction of comprehensive cancer information.²⁰ Separate from semantically extracting information from notes, other recent studies have focused on the use of aggregate data for outcome prediction.^{8,21}

Within symptom identification, work has been built off the larger field of adverse drug event monitoring. For cancer, this has been fairly limited; a prior work demonstrated the use of identifying topics within patient communications via the patient portal, demonstrating that side effect terms were associated with early discontinuation of hormonal therapy.²² This demonstrates an additional potential application of accurate symptom identification, as patients frequently will supplement PRO questionnaires with free text data.²³

This study is limited by its small sample of notes and expert reviewers at a single institution. Additionally, OTV notes are specifically intended to report toxicities and may overrepresent this information. The notes in our sample are also brief with very limited autopopulated text in comparison to documentation for other encounters. Thus, it is possible that the reported performance may not generalize across all oncology documentation. However, the gold standard data set upon which cTAKES was initially built was based on four total human annotators and 1556 annotations on 160 clinical notes,²⁴ and our study does serve as an external assessment for this specific use case. These limitations also underscore the labor intensity of expert review and emphasize the need for high-quality computational tools. Additionally, it is likely that our current pipeline, built from currently freely available “out-of-the-box” tools developed at a separate institution, would demonstrate additional accuracy with additional modifications.²⁵ Development of a separate model was considered, but it was ultimately decided to dedicate the statistical power of our manual annotation towards externally assessing the default configuration of a broadly available and used software package. Finally, this study focused on the extraction of symptoms in isolated notes across distinct patients. Attribution of toxicities would require temporal assessment across many notes, which would require more intensive abstraction by our expert reviewers. This study offers important use-case specific modifications and an independent external assessment of the tool. Importantly, we demonstrate that NLP offers good performance for the identification of specific symptoms in comparison to human expert reviewers.

While altering physician workflows to consistently prospectively document acute toxicities may be a potential option (adopted at some institutions), prior data suggest that it may underreport symptoms documented in the clinical chart.¹ Furthermore, the tools we evaluated and generated in this study are freely available online. NLP did not have strong detection of expert-identified negated symptoms, which likely reflects its greater complexity; expert-defined negations also identified specific scenarios that prompt disagreement during manual review.²

The implementation of NLP, in addition to offering an alternative to human review for the ascertainment of toxicities, can also be implemented to validate manually collected toxicities; this may augment the detection of toxicities on clinical trials and in retrospective research. Accurate extraction of clinical elements from the free text

may offer refined and rational features for predictive models, building on studies where aggregate text provided utility in predicting clinical outcomes.^{8,21} Our team recently completed one of the first prospective, randomized studies of machine learning, utilizing EHR data to generate accurate predictions of acute care, and direct supportive care.¹² NLP offers an additional source of insights from routine clinical data that may augment its performance for clinical decision support.^{11,26}

CONCLUSIONS

In conclusion, the use of an NLP pipeline facilitates CTCAE symptom identification via publicly available tools. In light of reviewer variability, this may serve as a tool to improve the consistency of toxicity capture in retrospective and prospective settings.

SUPPLEMENTARY MATERIAL

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

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. Publication made possible in part by support from the UCSF Open Access Publishing Fund.

CONFLICT OF INTEREST STATEMENT

J.D.T., M.P., and J.C.H. are coinventors on a pending patent, “Systems and methods for predicting acute care visits during outpatient cancer therapy,” broadly related to this manuscript.

DATA AVAILABILITY STATEMENT

The data underlying this article cannot be shared publicly due to the privacy of individuals whose data were used in the study.

REFERENCES

- Miller TP, Li Y, Kavcic M, *et al.* Accuracy of adverse event ascertainment in clinical trials for pediatric acute myeloid leukemia. *J Clin Oncol* 2016; 34 (13): 1537–43.
- Fairchild AT, Tanksley JP, Tenenbaum JD, *et al.* Inter-rater reliability in toxicity identification: limitations of current standards. *Int J Radiat Oncol Biol Phys* 2020; 107 (5): 996–1000.
- Miller TP, Fisher BT, Getz KD, *et al.* Unintended consequences of evolution of the common terminology criteria for adverse events. *Pediatr Blood Cancer* 2019; 66 (7): e27747.
- Rosenbloom ST, Denny JC, Xu H, *et al.* Data from clinical notes: a perspective on the tension between structure and flexible documentation. *J Am Med Inform Assoc* 2011; 18 (2): 181–6.
- 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* 2010; 17 (5): 507–13.
- Aronson AR. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. *Proc AMIA Symp* 2001; 17–21.
- Soysal E, Wang J, Jiang M, *et al.* CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J Am Med Inform Assoc* 2018; 25 (3): 331–6.
- Kehl KL, Elmarakeby H, Nishino M, *et al.* Assessment of deep natural language processing in ascertaining oncologic outcomes from radiology reports. *JAMA Oncol* 2019; 5 (10): 1421.
- Hripcsak G, Austin JHM, Alderson PO, *et al.* Use of natural language processing to translate clinical information from a database of 889,921 chest radiographic reports. *Radiology* 2002; 224 (1): 157–63.
- Xu H, Anderson K, Grann VR, *et al.* Facilitating cancer research using natural language processing of pathology reports. *Stud Health Technol Inform* 2004; 107 (Pt 1): 565–72.
- Hong JC, Niedzwiecki D, Palta M, *et al.* Predicting emergency visits and hospital admissions during radiation and chemoradiation: an internally validated pretreatment machine learning algorithm. *JCO Clin Cancer Inform* 2018; 2 (2): 1–11.
- Hong JC, Eclov NCW, Dalal NH, *et al.* System for High-Intensity Evaluation During Radiation Therapy (SHIELD-RT): A Prospective Randomized Study of Machine Learning-Directed Clinical Evaluations During Radiation and Chemoradiation. *JCO* 2020; 38 (31): 3652–61.
- Common Terminology Criteria for Adverse Events (CTCAE) | Protocol Development | CTEP. https://ctep.cancer.gov/protocolDevelopment/electronic_applications/ctc.htm (accessed September 19, 2019).
- Hong J. julianhong/ctcae. 2020. <https://github.com/julianhong/ctcae> (accessed October 16, 2020).
- Default Clinical Pipeline—Apache cTAKES—Apache Software Foundation. <https://cwiki.apache.org/confluence/display/CTAKES/Default+Clinical+Pipeline> (accessed October 10, 2020).
- Hripcsak G, Rothschild AS. Agreement, the F-measure, and reliability in information retrieval. *J Am Med Inform Assoc* 2005; 12 (3): 296–8.
- Revelle W. PSYCH: Procedures for Psychological, Psychometric, and Personality Research. 2020; <https://CRAN.R-project.org/package=psych> (accessed April 13, 2020).
- Gamer M, Lemon J, Singh IFP. irr: Various Coefficients of Interrater Reliability and Agreement; 2019. <https://CRAN.R-project.org/package=irr> (accessed April 13, 2020).
- 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.
- Savova GK, Tseytlin E, Finan S, *et al.* DeepPhe: a natural language processing system for extracting cancer phenotypes from clinical records. *Cancer Res* 2017; 77 (21): e115–8–e118.
- Gensheimer MF, Henry AS, Wood DJ, *et al.* Automated survival prediction in metastatic cancer patients using high-dimensional electronic medical record data. *J Natl Cancer Inst* 2019; 111 (6): 568–74.
- Yin Z, Harrell M, Warner JL, *et al.* The therapy is making me sick: how online portal communications between breast cancer patients and physicians indicate medication discontinuation. *J Am Med Inform Assoc* 2018; 25 (11): 1444–51.
- Chung AE, Shoenbill K, Mitchell SA, *et al.* Patient free text reporting of symptomatic adverse events in cancer clinical research using the National Cancer Institute’s Patient-Reported Outcomes version of the Common Terminology Criteria for Adverse Events (PRO-CTCAE). *J Am Med Inform Assoc* 2019; 26 (4): 276–85.
- Ogren P, Savova G, Chute C. Constructing evaluation corpora for automated clinical named entity recognition. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC’08). Marrakech, Morocco: European Language Resources Association (ELRA); May 28–30, 2008, 3143–50. http://www.lrec-conf.org/proceedings/lrec2008/pdf/796_paper.pdf (accessed August 12, 2020).
- Miller T, Geva A, Dligach D. Extracting adverse drug event information with minimal engineering. In: Proceedings of the 2nd Clinical Natural Language Processing Workshop. Minneapolis, Minnesota, USA: Association for Computational Linguistics; 2019, 22–7.
- Demner-Fushman D, Chapman WW, McDonald CJ. What can natural language processing do for clinical decision support? *J Biomed Inform* 2009; 42 (5): 760–72.