Original Paper

Natural Language Processing for Surveillance of Cervical and Anal Cancer and Precancer: Algorithm Development and Split-Validation Study

Carlos R Oliveira[1], MD, PhD; Patrick Niccolai[1]; Anette Michelle Ortiz[1], BSc; Sangini S Sheth[2], MD, MPH; Eugene D Shapiro[1,3], MD; Linda M Niccolai[3], PhD; Cynthia A Brandt[4,5], MD, MPH

[1]Department of Pediatrics, Yale University School of Medicine, New Haven, CT, United States

[2]Department of Obstetrics, Gynecology, and Reproductive Sciences, Yale University School of Medicine, New Haven, CT, United States

[3]Departments of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, United States

[4]Departments of Emergency Medicine, Biostatistics, and Health Informatics, Yale Schools of Medicine and Public Health, New Haven, CT, United States

[5]Veteran Affairs Connecticut Healthcare System, West Haven, CT, United States

Corresponding Author:
Carlos R Oliveira, MD, PhD
Department of Pediatrics
Yale University School of Medicine
P.O. Box 208000
New Haven, CT, 06520
United States
Phone: 1 203 785 5474
Email: carlos.oliveira@yale.edu

Abstract

Background: Accurate identification of new diagnoses of human papillomavirus–associated cancers and precancers is an important step toward the development of strategies that optimize the use of human papillomavirus vaccines. The diagnosis of human papillomavirus cancers hinges on a histopathologic report, which is typically stored in electronic medical records as free-form, or unstructured, narrative text. Previous efforts to perform surveillance for human papillomavirus cancers have relied on the manual review of pathology reports to extract diagnostic information, a process that is both labor- and resource-intensive. Natural language processing can be used to automate the structuring and extraction of clinical data from unstructured narrative text in medical records and may provide a practical and effective method for identifying patients with vaccine-preventable human papillomavirus disease for surveillance and research.

Objective: This study's objective was to develop and assess the accuracy of a natural language processing algorithm for the identification of individuals with cancer or precancer of the cervix and anus.

Methods: A pipeline-based natural language processing algorithm was developed, which incorporated machine learning and rule-based methods to extract diagnostic elements from the narrative pathology reports. To test the algorithm's classification accuracy, we used a split-validation study design. Full-length cervical and anal pathology reports were randomly selected from 4 clinical pathology laboratories. Two study team members, blinded to the classifications produced by the natural language processing algorithm, manually and independently reviewed all reports and classified them at the document level according to 2 domains (diagnosis and human papillomavirus testing results). Using the manual review as the gold standard, the algorithm's performance was evaluated using standard measurements of accuracy, recall, precision, and F-measure.

Results: The natural language processing algorithm's performance was validated on 949 pathology reports. The algorithm demonstrated accurate identification of abnormal cytology, histology, and positive human papillomavirus tests with accuracies greater than 0.91. Precision was lowest for anal histology reports (0.87, 95% CI 0.59-0.98) and highest for cervical cytology (0.98, 95% CI 0.95-0.99). The natural language processing algorithm missed 2 out of the 15 abnormal anal histology reports, which led to a relatively low recall (0.68, 95% CI 0.43-0.87).

Conclusions: This study outlines the development and validation of a freely available and easily implementable natural language processing algorithm that can automate the extraction and classification of clinical data from cervical and anal cytology and histology.

(JMIR Med Inform 2020;8(11):e20826) doi: 10.2196/20826

KEYWORDS

natural language processing; automated data extraction; human papillomavirus; surveillance; pathology reporting; cervical cancer; anal cancer; precancer; cancer; HPV; accuracy

Introduction

Precision public health is a rapidly evolving field that focuses on promoting the health of a population through the application of technology [1]. A key priority in precision public health is the development of new informatics approaches to optimize the use of vaccines for the prevention of disease. Some of the more successful vaccine informatics applications postlicensure include using text-mining techniques to automate the tracking of adverse immunization outcomes and the use of emergency department notes as an early warning sign for outbreaks of vaccine-preventable diseases. Automation of biosurveillance and timely identification of infectious diseases is of particular importance to public health, as it allows for better planning and distribution of limited resources [2-4].

Persistent infection with human papillomavirus (HPV) can result in precancerous anogenital lesions as well as invasive cancer. In the United States, approximately 25,000 cases of anogenital cancers are diagnosed every year, with cervical and anal cancer being the majority (75%) of these [5]. Over 90% of these cases are attributable to infection with HPV types that are preventable by the use of recommended HPV vaccines [5-7]. Although HPV vaccines have high proven efficacy, the way we use these vaccines to prevent HPV cancers is still in need of improvement [8]. Accurate identification and tracking of new cases of HPV cancers is an important step toward the development of strategies that optimize the use of HPV vaccines.

Surveillance for HPV-associated outcomes is critical for monitoring the progress of immunization programs and identifying targets for improvement. Surveillance for HPV cancers, however, has been a formidable challenge. Most of the clinical data needed to diagnose a patient with an HPV-related cancer, or precancer, are stored in pathology reports. Normally, pathology reports are stored in a narrative format and contain several lines of text that can include nondiagnostic information, such as medical history or clinical indications for screening [9]. Although a manual review of these free-text pathology reports is the most accurate case-finding method, it is a laborious process that can become too impractical for large-scale surveillance projects. To facilitate data capture and analysis, considerable efforts have been made to promote processes that encourage pathologists to document their findings in a specific format and using standardized terminology [10]. However, most efforts to incorporate standardized reporting have yet to be consistently implemented by health care providers and institutions [11].

To develop an accurate and scalable surveillance platform for HPV vaccine-preventable cancers, it is critical to first overcome the challenge of narrative data-abstraction. A potential solution to this data-abstraction problem is automation with computational tools, such as natural language processing (NLP). NLP is an increasingly used approach that combines informatics and linguistic techniques to automatically identify and extract key concepts or phrases embedded in a narrative text [12]. Although NLP has been successfully applied for the surveillance of several cancers (eg, colon, hepatic, and bladder cancer), it has been underutilized for the surveillance of HPV cancers and precancers [12-15].

As a first step toward achieving automated surveillance of HPV vaccine-preventable diseases, we developed an NLP algorithm aiming to extract information from cervical and anal pathology reports and classify these reports based on the pathologist's final diagnosis. The objective of this study was to assess the accuracy of our NLP algorithm for the identification of individuals with cancer or precancer of the cervix and anus.

Methods

Study Design and Setting

This study used data generated from the HPV Vaccine Effectiveness Project, a large-scale population-based study aiming to determine the effectiveness of the HPV vaccine [16]. In support of this ongoing project, an NLP algorithm was developed to convert narrative pathology reports into structured data that can be queried to identify individuals who had HPV-related abnormalities in their cervical or anal pathology report. To build and evaluate this NLP algorithm, a split-validation method was used, wherein 2 sets of full-length cervical and anal pathology reports were randomly selected from 4 different clinical pathology laboratories within the Yale–New Haven Health System participating in the HPV Vaccine Effectiveness Project. The first set of reports was used to build the algorithm (ie, the training set, n=100), and the second set was used for testing the accuracy of the algorithm (ie, the validation set, n=1000). Pathology reports were extracted between January 1, 2010 and December 31, 2018 and deidentified for both the development and testing phases of this study.

NLP Algorithm Development

We developed a pipeline-based NLP algorithm that incorporated both machine learning and rule-based methods to extract and classify diagnostic elements (histopathology, cytopathology, and HPV test results) from narrative pathology reports. Various software platforms have been developed to automatically

annotate and process clinical notes based on the Unstructured Information Management Architecture framework [17-19]. Our pipeline was built using CLAMP (Clinical Language Annotation, Modeling, and Processing) software, because it is open-source, modular, free-to-use, and specifically designed to process and analyze clinical text [20]. Our pipeline combined several existing and well-validated text processing components [21-27] and built on these components with newly developed HPV-specific ontologies and postprocessing features.

NLP Data Extraction

The first steps of our pipeline involved using CLAMP's existing algorithms to preprocess each report and apply a series of if-then rules to parse and enumerate each sentence and word within the full-length report (ie, a sentence detector and word tokenizer, respectively) [24]. Next, we used a supervised machine learning approach to assign each enumerated token (ie, each word or set of words) a tag based on its part of speech (eg, verb, noun, etc) [28]. A more in-depth description of the pipeline's individual preprocessing components can be found in Multimedia Appendix 1. We then implemented an existing named entity recognizer program to identify key concepts within the narrative text [29]. This named entity recognizer program utilizes a dictionary-based approach to match concepts in pathology reports to terms in a dictionary derived from the Unified Medical Language System Metathesaurus [27]. To more robustly account for variations in HPV-related concepts, we also constructed an HPV-cancer dictionary and incorporated it into the algorithm. This custom HPV-cancer dictionary leveraged over a decade of experience and expertise in HPV-cancer surveillance through collaboration with seasoned epidemiologists from HPV Vaccine Impact Monitoring Project Across Connecticut, a collaborative project between the Connecticut Emerging Infections Program at Yale School of Public Health; the Connecticut Department of Public Health; and the Centers for Disease Control and Prevention [30]. We have contributed our HPV dictionary (ie, ontology) to the National Center for Biomedical Ontology BioPortal platform [31], where it is openly available for other users to develop further.

NLP Data Classification

After implementing the dictionary-based named entity recognizer, we applied newly developed heuristic rules to analyze and relabel each concept based on their context in the report. For example, a series of if-then rules were employed to identify different sections of the report (eg, clinical history, molecular diagnosis, primary diagnosis, etc) and determine when an HPV-related diagnosis was being stated in the report as a historical piece of information and when it was being stated in the context of the current specimen. Further details and examples of the key if-then rules are shown in Multimedia Appendix 1.

We also implemented an extensively validated rule-based negation algorithm [23] to allow us to differentiate when a recognized concept was being negated or stated with uncertainty based on the words that preceded or followed the identified concept (eg, "negative for abnormalities" or "abnormalities were not found"). Once all entities were named, coded, and contextualized, the algorithm generated a structured output (matrix) that was suitable for further processing. For the last step of the algorithm, the structured output was used to summarize and classify each report, at the document level, in 2 key domains: final diagnosis (using the Bethesda Classification system) and results of HPV tests (if performed). To enable the reproducibility of this study, our pipeline was freely available for research through CLAMP [32] and is archived [33]. To facilitate its application, we also provide a step-by-step video demonstration of this pipeline [33].

Classification Validation

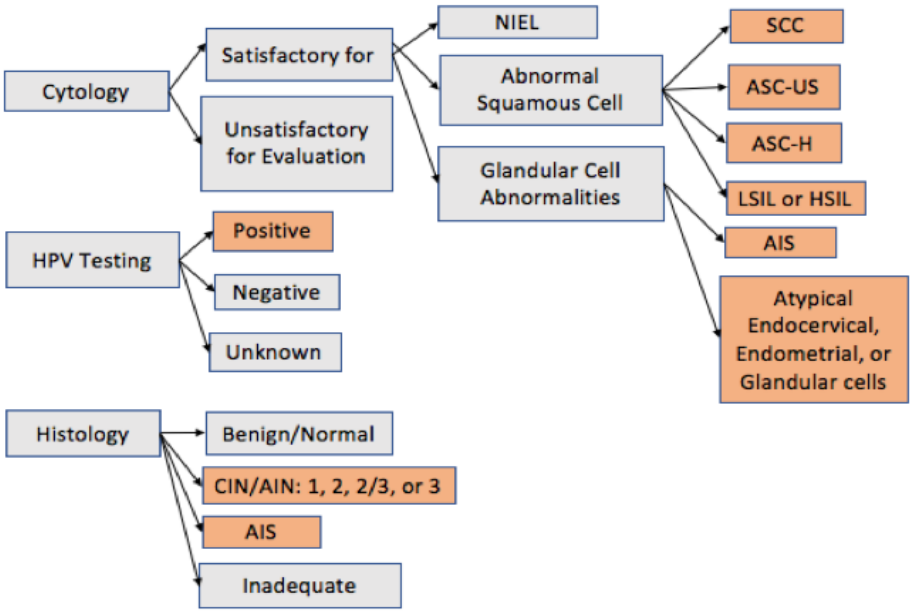
To test the algorithm's classification accuracy, 2 study team members, blinded to the classifications produced by the NLP algorithm, manually and independently reviewed all pathology reports in the validation set and classified them at the document level according to the same 2 domains (diagnosis and HPV testing results). Disagreement among the 2 manual-review adjudicators was resolved by discussion with a third investigator.

For the primary analysis, we tested this algorithm's accuracy for the identification of HPV-related pathology. The primary outcome—abnormal pathology—was grouped as a dichotomous variable and defined, for cytology reports, as a final diagnosis of atypical squamous cells or greater, and for histology reports, as intraepithelial neoplasia grades 2 or greater. A summary of the classification process for the primary outcome is shown in Figure 1.

Statistical Analysis

The algorithm's performance was evaluated using the manual review classifications as the standard. Accuracy, precision, recall, and F-measure were calculated as follows: $accuracy = (true\ positives + true\ negatives) / (true\ positives + true\ negatives + false\ positives + false\ negatives)$; $precision = true\ positives / (true\ positives + false\ positives)$; $recall = true\ positives / (true\ positives + false\ negatives)$; $F\text{-}measure = 2 \times (precision \times recall) / (precision + recall)$. Statistical analyses were conducted using Stata statistical software (version 15; StataCorp LLC). This protocol was approved by the institutional review board of Yale University (protocol number 2000024708).

Figure 1. Diagrammatic representation of the classification process for pathology reports (colored indicates abnormal pathology). AIN: anal intraepithelial lesion; AIS: adenocarcinoma in situ; ASC-US: atypical squamous cells of undetermined significance; ASC-H: atypical squamous cells—cannot exclude high-grade squamous intraepithelial lesion; CIN: cervical intraepithelial lesion;HSIL: high-grade squamous intraepithelial lesion; LSIL: low-grade squamous intraepithelial lesion; NIEL: negative for intraepithelial lesion; SCC: squamous cell carcinoma.



Results

Out of 1000 pathology reports originally selected for the validation set, 51 were excluded after manual review because they were (1) reports with misclassified specimens (ie, not anal or cervical tissue), (2) duplicate reports, or (3) incomplete reports. Testing of the NLP algorithm's accuracy was performed on 949 pathology reports (anal cytology n=94; anal histology n=86; cervical cytology n=403; cervical histology n=366). HPV tests were documented on 303 reports (cervical n=265; anal cytology n=38), of which 121 (40%) had positive results for HPV. A summary of the highest-grade diagnosis based on manual review of the 949 pathology reports is shown in Table 1. Most of the biopsies performed revealed either normal or low-grade (362/452, 80%) lesions, and most of the cytologic specimens were negative for intraepithelial lesions (302/497, 61%).

Table 1. Summary of results from the manual review of the validation set.

Test	Cervical (n=769), n (%)	Anal (n=180), n (%)	Total (N=949), n
Cytology	403 (81.1)	94 (18.9)	497
Negative for intraepithelial lesion	255 (84.4)	47 (15.6)	302
Atypical squamous cells of undetermined significance	44 (68.8)	20 (31.3)	64
Atypical squamous cells—cannot exclude high-grade squamous intraepithelial lesion	57 (98.3)	1 (1.7)	58
Low-grade squamous intraepithelial lesion	16 (84.2)	3 (15.8)	19
Glandular abnormality	14 (82.4)	3 (17.6)	17
Unsatisfactory specimen	17 (45.9)	20 (54.1)	37
HPV[a] test performed	206 (85.8)	34 (14.2)	240
Positive	91 (84.3)	17 (15.7)	108
Histology	366 (81.0)	86 (19.0)	452
Benign	153 (77.3)	45 (22.7)	198
Squamous intraepithelial lesion grade 1	138 (84.1)	26 (15.9)	164
Squamous intraepithelial lesion grade 2+	75 (83.3)	15 (16.7)	90

[a]HPV: human papillomavirus.

For the primary analysis, the NLP algorithm accurately identified abnormal cytology, histology, and positive HPV tests with accuracies ≥0.91 in all specimens (Table 2). Precision was lowest for anal histology reports (0.87, 95% CI 0.59-0.98) and

highest for cervical cytology (0.98, 95% CI 0.95-0.99). The NLP algorithm missed 2 out of the 15 abnormal anal histology reports, which led to relatively low recall (0.68, 95% CI 0.43-0.87).

Table 2. Performance of NLP algorithm on the validation set, N = 949.

Variable	Precision (95% CI)	Recall (95% CI)	F-measure (95% CI)	Accuracy (95% CI)
Abnormal cytology[a]				
Cervical	0.98 (0.95-0.99)	1.00 (0.97-1.00)	0.99 (0.98-1.00)	0.99 (0.98-1.00)
Anal	0.93 (0.76-0.99)	1.00 (0.86-1.00)	0.96 (0.91-1.00)	0.98 (0.93-0.99)
HPV[b] **testing**				
Positive	0.95 (0.89-0.98)	1.00 (0.97-1.00)	0.97 (0.95-0.99)	0.99 (0.98-1.00)
Abnormal histology				
CIN[c] grade 2+	0.89 (0.80-0.95)	0.93 (0.85-0.98)	0.91 (0.86-0.96)	0.96 (0.94-0.98)
AIN[d] grade 2+	0.87 (0.59-0.98)	0.68 (0.43-0.87)	0.76 (0.61-0.92)	0.91 (0.82-0.96)
Average performance[e]				
Abnormal test	0.94 (0.91-0.97)	0.96 (0.92-0.98)	0.94 (0.93-0.97)	0.97 (0.96-0.98)

[a]Abnormalities include atypical squamous cells of undetermined significance, atypical squamous cells—cannot exclude high-grade squamous intraepithelial lesion, low-grade squamous intraepithelial lesion, and glandular cell abnormalities.

[b]HPV: human papillomavirus.

[c]CIN: cervical intraepithelial lesion.

[d]AIN: anal intraepithelial lesion.

[e]Includes results from both cytology and histology.

Discussion

In this paper, we described the development and validation of an NLP instrument that can be used for both data extraction and classification of cytology and histology reports of the cervix and anus. Based on these initial data, our NLP algorithm can classify whether a cytology or histology specimen was abnormal and whether any HPV tests resulted positive, with an accuracy 91%. At the document level, this algorithm had an average recall (also known as sensitivity) of 96% and precision (also known as positive predictive value) of 94%. This demonstration of accuracy is an important first step toward the development of a tool that can facilitate the automation of surveillance for HPV vaccine-preventable cancers and precancers.

There is an increasing body of evidence showing the merits of an NLP system over manual review for data extraction and document classification for disease surveillance [34,35]. A key contribution of this study is the integration and application of well-validated NLP methodologies to solve a real-world public health problem. Most individual components included in our NLP pipeline have been previously validated. Using a commonly used corpus (SemEval-2014), Soysal et al [20] demonstrated that CLAMP's named entity recognizer algorithm had superior precision to those of other commonly used platforms (CLAMP: 0.77; MetaMa*P*: 0.55; cTAKES: 0.46). In the same study [20], the performance accuracy of other key components (tokenizer, sentence boundary detector, part-of-speech tagger, and section detector) were evaluated using the MiPACQ clinical corpus and were also found to have a high accuracy (>92%). In our study, we did not aim to develop novel NLP strategies or components. However, one of the key strengths in our approach is that we were able to leverage the experience of HPV surveillance experts to assemble an extensive list of HPV-related terms to optimize named entity recognition.

This study has several other notable strengths. First, this study is among the first to evaluate the accuracy of an NLP algorithm to identify cases of HPV-related precancers. Although precancerous diagnoses are routinely made, these data are not systematically collected by most surveillance systems. These diagnoses, however, have public health significance as they can be used to monitor the impact of HPV vaccines. Our NLP algorithm provides an efficient way to use existing resources to measure the extent to which HPV vaccines reduce the burden of disease at the population level and identify areas to strengthen immunization programs. Automating the identification of precancers may also have clinical applications. For example, following an abnormal cytology result, a patient is usually kept under close surveillance for months. After an abnormal cytology screen, the appropriate management can vary from more frequent follow-up tests to immediate treatment with surgical excision. Automation of the detection of precancerous abnormalities in cytology or histology can be incorporated into clinical decision support tools to ensure patients are appropriately linked to care and are receiving timely follow-up.

An additional strength of this study is in the application of our NLP algorithm to accurately detect cases of anal cancer and precancer. To our knowledge, we are the first to provide a tool specifically designed for this purpose. Efforts to monitor the impact of HPV vaccination on oncogenic outcomes have focused mainly on cervical cancer and women. With the increased

recognition that HPV also causes cancer in men and the increasing rate of these cancers in the young adult population [25,26], it is important to determine if the HPV vaccine's deployment can be optimized to reduce the burden of disease in both sexes. A surveillance system with these outcomes may be especially valuable to investigators and public health officials in assessing the impact of various immunization strategies in both males and females.

Additional improvements can optimize the performance of this algorithm for implementation in routine public health surveillance or clinical practice. For example, we only used reports from a single health care system (Yale New Haven Health), which likely limited the variability found in both the structure and language in the pathology reports. Thus, future work is needed to validate this tool's portability to other health care systems where pathology practices may differ. An additional area of improvement is in the preprocessing. After initial manual review of pathology reports, we had to exclude several reports that were incomplete or were misclassified in the electronic medical record. To be useful as a real-time surveillance tool, future iterations of this NLP algorithm will need to address the potential for misclassification at the onset. An additional limitation of this tool is that it was developed as a means to identify cases of cancer and precancer at the document level and not at the patient level. As many individuals have more than one pathology report in their record, to be useful as an automated surveillance method, more postprocessing will be needed to deal with duplicates or disparate findings at the patient level.

In this study, we detail the development of a freely available and easily implementable NLP algorithm that can automate the extraction of clinical data from cervical and anal cytology and histology reports. We show that with this algorithm, it is possible to accurately detect patients with HPV-related abnormalities at these anatomical sites. These data provide preliminary support for the use of our NLP instrument for the surveillance of HPV cancer and precancer of the cervix and anus.

Acknowledgments

We would like to acknowledge the team of investigators and epidemiologists at HPV Vaccine Impact Monitoring Project Across Connecticut: Kyle Higgins, Monica Brackney, and James Meek.

This work was supported in part by National Institutes of Health grant number R01AI123204 (PN) from the National Institute of Allergy and Infectious Diseases and grant numbers KL2TR001862 (CRO) and UL1TR000142 (EDS) from the National Center for Advancing Translational Science. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of National Institutes of Health. The funders of the study had no role in study design, data collection, data analysis, data interpretation, or writing of the report. The corresponding author had full access to all of the data in this study and had final responsibility for the decision to submit for publication.

Conflicts of Interest

LMN reports previous work as a scientific advisor for Merck. SSS has previously provided consulting services to Merck and received a research grant from Merck. All other authors declare no conflicts of interest.

Multimedia Appendix 1

Supplementary methods.
[DOCX File , 894 KB-Multimedia Appendix 1]

References

1. Bayer R, Galea S. Public health in the precision-medicine era. N Engl J Med 2015 Aug 06;373(6):499-501. [doi: 10.1056/NEJMp1506241] [Medline: 26244305]
2. Yu W, Zheng C, Xie F, Chen W, Mercado C, Sy LS, et al. The use of natural language processing to identify vaccine-related anaphylaxis at five health care systems in the Vaccine Safety Datalink. Pharmacoepidemiol Drug Saf 2020 Feb;29(2):182-188. [doi: 10.1002/pds.4919] [Medline: 31797475]
3. Elkin PL, Froehling DA, Wahner-Roedler DL, Brown SH, Bailey KR. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. Ann Intern Med 2012 Jan 03;156(1 Pt 1):11-18. [doi: 10.7326/0003-4819-156-1-201201030-00003] [Medline: 22213490]
4. Ye Y, Tsui FR, Wagner M, Espino JU, Li Q. Influenza detection from emergency department reports using natural language processing and Bayesian network classifiers. J Am Med Inform Assoc 2014;21(5):815-823 [FREE Full text] [doi: 10.1136/amiajnl-2013-001934] [Medline: 24406261]
5. Centers for Disease Control and Prevention. Cancers Associated with Human Papillomavirus, United States—2012–2016. USCS Data Brief.: US Department of Health and Human Services; 2019. URL: www.cdc.gov/cancer/uscs/about/data-briefs/no10-hpv-assoc-cancers-UnitedStates-2012-2016.htm [accessed 2020-10-27]
6. Petrosky E, Bocchini JA, Hariri S, Chesson H, Curtis CR, Saraiya M, Centers for Disease Control and Prevention. Use of 9-valent human papillomavirus (HPV) vaccine: updated HPV vaccination recommendations of the advisory committee on immunization practices. MMWR Morb Mortal Wkly Rep 2015 Mar 27;64(11):300-304 [FREE Full text] [Medline: 25811679]

7. Gargano J, Meites E, Watson M, Unger E, Markowitz L. Chapter 5: human papillomavirus. In: Roush SW, Baldy LM, Kirkconnell Hall MA, editors. Manual for the Surveillance of Vaccine-Preventable Diseases. Atlanta, GA: Centers for Disease Control and Prevention Department of Health and Human Services; Apr 28, 2020.
8. Sivaram S, Sanchez MA, Rimer BK, Samet JM, Glasgow RE. Implementation science in cancer prevention and control: a framework for research and programs in low- and middle-income countries. Cancer Epidemiol Biomarkers Prev 2014 Nov;23(11):2273-2284. [doi: 10.1158/1055-9965.EPI-14-0472] [Medline: 25178984]
9. Crothers BA, Tench WD, Schwartz MR, Bentz JS, Moriarty AT, Clayton AC, et al. Guidelines for the reporting of nongynecologic cytopathology specimens. Arch Pathol Lab Med 2009 Nov;133(11):1743-1756. [doi: 10.1043/1543-2165-133.11.1743] [Medline: 19886707]
10. Renshaw AA, Mena-Allauca M, Gould EW, Sirintrapun SJ. Synoptic reporting: evidence-based review and future directions. JCO Clin Cancer Inform 2018 Dec;2:1-9. [doi: 10.1200/CCI.17.00088] [Medline: 30652566]
11. Bodenreider O, Cornet R, Vreeman DJ. Recent developments in clinical terminologies - SNOMED CT, LOINC, and RxNorm. Yearb Med Inform 2018 Aug;27(1):129-139 [FREE Full text] [doi: 10.1055/s-0038-1667077] [Medline: 30157516]
12. Imler TD, Morea J, Kahi C, Imperiale TF. Natural language processing accurately categorizes findings from colonoscopy and pathology reports. Clin Gastroenterol Hepatol 2013 Jun;11(6):689-694 [FREE Full text] [doi: 10.1016/j.cgh.2012.11.035] [Medline: 23313839]
13. Wagholikar KB, MacLaughlin KL, Henry MR, Greenes RA, Hankey RA, Liu H, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc 2012;19(5):833-839. [doi: 10.1136/amiajnl-2012-000820] [Medline: 22542812]
14. Schroeck FR, Patterson OV, Alba PR, Pattison EA, Seigne JD, DuVall SL, et al. Development of a natural language processing engine to generate bladder cancer pathology data for health services research. Urology 2017 Dec;110:84-91 [FREE Full text] [doi: 10.1016/j.urology.2017.07.056] [Medline: 28916254]
15. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al. Natural language processing technologies in radiology research and clinical applications. Radiographics 2016;36(1):176-191 [FREE Full text] [doi: 10.1148/rg.2016150080] [Medline: 26761536]
16. Oliveira CR. Estimating the Effectiveness of Human Papillomavirus Vaccine: A Case-Control Study with Bayesian Model Averaging. In: Yale University. Ann Arbor: ProQuest Dissertations & Theses Global; 2019:126.
17. Bates J, Fodeh SJ, Brandt CA, Womack JA. Classification of radiology reports for falls in an HIV study cohort. J Am Med Inform Assoc 2016 Apr;23(e1):e113-e117 [FREE Full text] [doi: 10.1093/jamia/ocv155] [Medline: 26567329]
18. Garla V, Lo RV, Dorey-Stein Z, Kidwai F, Scotch M, Womack J, et al. The Yale cTAKES extensions for document classification: architecture and application. J Am Med Inform Assoc 2011;18(5):614-620. [doi: 10.1136/amiajnl-2011-000093] [Medline: 21622934]
19. Womack JA, Murphy TE, Rentsch CT, Tate JP, Bathulapalli H, Smith AC, et al. Polypharmacy, hazardous alcohol and illicit substance use, and serious falls among PLWH and uninfected comparators. J Acquir Immune Defic Syndr 2019 Nov 01;82(3):305-313. [doi: 10.1097/QAI.0000000000002130] [Medline: 31339866]
20. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc 2018 Mar 01;25(3):331-336. [doi: 10.1093/jamia/ocx132] [Medline: 29186491]
21. Le DV, Montgomery J, Kirkby KC, Scanlan J. Risk prediction using natural language processing of electronic mental health records in an inpatient forensic psychiatry setting. J Biomed Inform 2018 Oct;86:49-58 [FREE Full text] [doi: 10.1016/j.jbi.2018.08.007] [Medline: 30118855]
22. Redman JS, Natarajan Y, Hou JK, Wang J, Hanif M, Feng H, et al. Accurate identification of fatty liver disease in data warehouse utilizing natural language processing. Dig Dis Sci 2017 Oct;62(10):2713-2718. [doi: 10.1007/s10620-017-4721-9] [Medline: 28861720]
23. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform 2001 Oct;34(5):301-310 [FREE Full text] [doi: 10.1006/jbin.2001.1029] [Medline: 12123149]
24. Doan S, Bastarache L, Klimkowski S, Denny JC, Xu H. Integrating existing natural language processing tools for medication extraction from discharge summaries. J Am Med Inform Assoc 2010 Oct;17(5):528-531. [doi: 10.1136/jamia.2010.003855] [Medline: 20819857]
25. Brotherton JML, Giuliano AR, Markowitz LE, Dunne EF, Ogilvie GS. Monitoring the impact of HPV vaccine in males-considerations and challenges. Papillomavirus Res 2016 Dec;2:106-111 [FREE Full text] [doi: 10.1016/j.pvr.2016.05.001] [Medline: 29074169]
26. Palefsky JM. Human papillomavirus-related disease in men: not just a women's issue. J Adolesc Health 2010 Apr;46(4 Suppl):S12-S19 [FREE Full text] [doi: 10.1016/j.jadohealth.2010.01.010] [Medline: 20307839]
27. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res 2004 Jan 01;32(Supp 1):D267-D270 [FREE Full text] [doi: 10.1093/nar/gkh061] [Medline: 14681409]
28. Apache OpenNLP Development Community. Apache OpenNLP.: The Apache Software Foundation; 2020. URL: http://opennlp.apache.org/index.html [accessed 2020-10-27]

29. Uzuner O, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. J Am Med Inform Assoc 2011;18(5):552-556 [FREE Full text] [doi: 10.1136/amiajnl-2011-000203] [Medline: 21685143]
30. Hariri S, Markowitz LE, Bennett NM, Niccolai LM, Schafer S, Bloch K, Hpv-Impact Working Group. Monitoring effect of human papillomavirus vaccines in US population, emerging infections program, 2008-2012. Emerg Infect Dis 2015 Sep;21(9):1557-1561 [FREE Full text] [doi: 10.3201/eid2109.141841] [Medline: 26291379]
31. National Center for Biomedical Ontology. Bioportal.: The Board of Trustees of Leland Stanford Junior University; 2020. URL: http://bioportal.bioontology.org/ontologies/HPV [accessed 2020-10-27]
32. Hao D. Clinical Language Annotation, Modeling, and Processing Toolkit. Houston, TX: Center for Computational Biomedicine; 2020. URL: http://clamp.uth.edu [accessed 2020-10-27]
33. Niccolai P, Oliveira CR. HPV Pathology CLAMP Pipeline.: GitHub; 2020. URL: https://github.com/PatrickNiccolai1/HPV_Pathology_Clamp_Pipeline [accessed 2020-10-27]
34. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, Jones SF, et al. Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. J Biomed Inform 2017 Dec;73:14-29 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.012] [Medline: 28729030]
35. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. J Am Med Inform Assoc 2016 Nov;23(6):1077-1084. [doi: 10.1093/jamia/ocw006] [Medline: 27026618]

Abbreviations

CLAMP: Clinical Language Annotation, Modeling, and Processing
HPV: human papillomavirus
NLP: natural language processing

Edited by G Eysenbach; submitted 29.05.20; peer-reviewed by S Noah, S Doan, Y Motoki; comments to author 19.06.20; revised version received 18.09.20; accepted 04.10.20; published 03.11.20

Please cite as:
Oliveira CR, Niccolai P, Ortiz AM, Sheth SS, Shapiro ED, Niccolai LM, Brandt CA
Natural Language Processing for Surveillance of Cervical and Anal Cancer and Precancer: Algorithm Development and Split-Validation Study
JMIR Med Inform 2020;8(11):e20826
URL: https://medinform.jmir.org/2020/11/e20826
doi: 10.2196/20826
PMID: 32469840