JAMA Otolaryngology–Head & Neck Surgery | Original Investigation | FROM THE AMERICAN HEAD AND NECK SOCIETY

Oropharyngeal Cancer Staging Health Record Extraction Using Artificial Intelligence

Elif Baran, HBSc; Melissa Lee, MD; Steven Aviv, BBusSc; Jessica Weiss, MSc; Chris Pettengell, MD; Irene Karam, MD; Andrew Bayley, MD; Ian Poon, MD; Kelvin K. W. Chan, MD, PhD; Ambica Parmar, MD; Martin Smoragiewicz, MD, CM, PhD; Hagen Klieb, DMD; Tra Truong, MD; Pejman Maralani, MD; Danny J. Enepekides, MD; Kevin M. Higgins, MD; Antoine Eskander, MD

IMPORTANCE Accurate, timely, and cost-effective methods for staging oropharyngeal cancers are crucial for patient prognosis and treatment decisions, but staging documentation is often inaccurate or incomplete. With the emergence of artificial intelligence in medicine, data abstraction may be associated with reduced costs but increased efficiency and accuracy of cancer staging.

OBJECTIVE To evaluate an algorithm using an artificial intelligence engine capable of extracting essential information from medical records of patients with oropharyngeal cancer and assigning tumor, nodal, and metastatic stages according to American Joint Committee on Cancer eighth edition guidelines.

DESIGN, SETTING, AND PARTICIPANTS This retrospective diagnostic study was conducted among a convenience sample of 806 patients with oropharyngeal squamous cell carcinoma. Medical records of patients with staged oropharyngeal squamous cell carcinomas who presented to a single tertiary care center between January 1, 2010, and August 1, 2020, were reviewed. A ground truth cancer stage dataset and comprehensive staging rule book consisting of 135 rules encompassing p16 status, tumor, and nodal and metastatic stage were developed. Subsequently, 4 distinct models were trained: model T (entity relationship extraction) for anatomical location and invasion state, model S (numerical extraction) for lesion size, model M (sequential classification) for metastasis detection, and a p16 model for p16 status. For validation, results were compared against ground truth established by expert reviewers, and accuracy was reported. Data were analyzed from March to November 2023.

MAIN OUTCOMES AND MEASURES The accuracy of algorithm cancer stages was compared with ground truth.

RESULTS Among 806 patients with oropharyngeal cancer (mean [SD] age, 63.6 [10.6] years; 651 males [80.8%]), 421 patients (52.2%) were positive for human papillomavirus. The artificial intelligence engine achieved accuracies of 55.9% (95% CI, 52.5%-59.3%) for tumor, 56.0% (95% CI, 52.5%-59.4%) for nodal, and 87.6% (95% CI, 85.1%-89.7%) for metastatic stages and 92.1% (95% CI, 88.5%-94.6%) for p16 status. Differentiation between localized (stages 1-2) and advanced (stages 3-4) cancers achieved 80.7% (95% CI, 77.8%-83.2%) accuracy.

CONCLUSION AND RELEVANCE This study found that tumor and nodal staging accuracies were fair to good and excellent for metastatic stage and p16 status, with clinical relevance in assigning optimal treatment and reducing toxic effect exposures. Further model refinement and external validation with electronic health records at different institutions are necessary to improve algorithm accuracy and clinical applicability.

JAMA Otolaryngol Head Neck Surg. doi:10.1001/jamaoto.2024.1201
Published online May 16, 2024.

Supplemental content

Author Affiliations: Author affiliations are listed at the end of this article.

Corresponding Author: Antoine Eskander, MD, Department of Otolaryngology–Head and Neck Surgery, Sunnybrook Health Sciences Centre, University of Toronto, 2075 Bayview Ave, Room M1 102, Toronto, ON M4N 3M5, Canada (antoine.eskander@mail.utoronto.ca).

Cancer staging is crucial to acknowledging disease extent and providing optimal patient treatment, yet data pertaining to cancer stages are noted to be inaccurate or incomplete.[1,2] Although cancer centers have trained experts who document cancer staging to maintain in records, this process exhausts human resources because data abstraction is time-consuming and expensive.[3] Additionally, staging data provided do not offer much use in treatment planning given that data abstraction takes place months after the commencement of treatments and is often never entered into the medical record.[3] One of the hopes in implementing artificial intelligence (AI) in medicine is to reduce costs and effort related to data abstraction in an accurate and efficient matter while not burdening health care practitioners with additional work.[4]

Natural language processing, a type of machine learning AI, was previously demonstrated to be successful in extracting key information in reports, with overall success rates nearing 99%.[5] Other studies have been successful in determining tumor (T), nodal (N), and metastases (M) stages in lung cancer with moderate to high levels of accuracy from synoptic pathology reports.[1,2] This study focused on the development of a machine learning algorithm to automate oropharyngeal squamous cell carcinoma (OPSCC) cancer staging from unstructured clinical documentation. Early stages of OPSCC are commonly treated with radiotherapy, while locally advanced stages are treated with chemoradiotherapy.[6] It is therefore crucial to develop a novel algorithm capable of capturing staging-related data from clinical and radiology reports given that patients with OPSCC rarely undergo resection at this study's center and thus lack sufficient pathology reports to determine disease extent. Ultimately, however, human papillomavirus (HPV) status in OPSCC changes the prognosis and treatment type (ie, a deescalated treatment for patients who are p16 positive compared with that of patients who are p16 negative), and documentation is a quality metric.[7] As a result, information from pathology reports must be extracted to determine HPV status using p16 biomarkers.

The objective of this study was to develop and evaluate a novel approach using AI engine Darwen (Pentavere) (using natural language processing) that was capable of extracting staging information from clinical, radiology, and pathology reports in electronic health records (EHRs) of patients with OPSCC and automate TNM staging according to the American Joint Committee on Cancer (AJCC) eighth edition cancer staging guidelines. For the development of this novel approach, a TNM-labeled ground truth dataset and an additional, text-based rule book were used.

Key Points

Question Can artificial intelligence accurately stage oropharyngeal squamous cell carcinomas from medical records?

Findings In this diagnostic study among 806 patients with oropharyngeal cancer, artificial intelligence had low accuracy for classification of tumor and nodal cancer stages. Binary outputs of metastatic stage and p16 status had the highest accuracy, and overall accuracy in distinguishing localized (stages 1-2) vs advanced cancer (stages 3-4) was high.

Meaning These results suggest that artificial intelligence may be associated with enhanced patient care and oncological decision-making in patients with oropharyngeal squamous cell carcinoma through detection of localized vs advanced cancer stages.

Methods

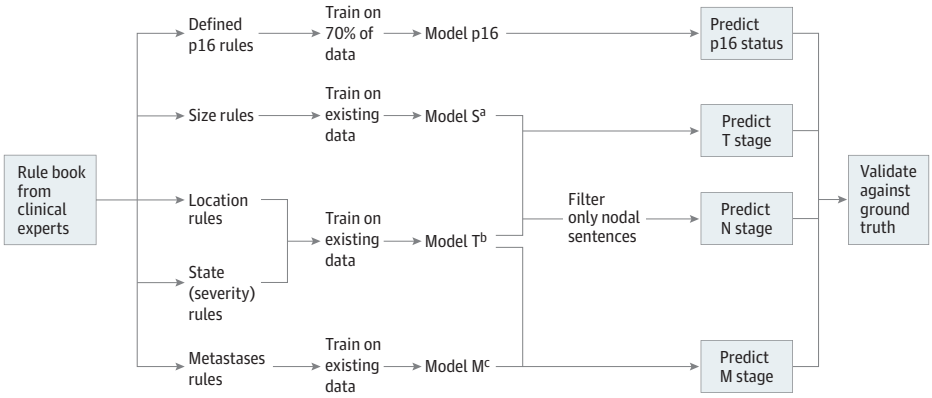
In this diagnostic study, clinical, radiology, and pathology reports in EHRs of 806 patients with oropharyngeal cancer seen at the Odette Cancer Center (OCC) at Sunnybrook Health Sciences Center between January 1, 2010, and August 1, 2020, were manually reviewed by 2 expert reviewers (E.B. and A.E.). Each patient was reviewed to collect information pertaining to p16 status and TNM and overall cancer stages to determine clinical stages based on the AJCC 8th edition cancer staging guidelines.[8] The staging provided by expert reviewers (E.B. and A.E.) was labeled ground truth. The study protocol was approved by the research ethics board of Sunnybrook Health Sciences Center. Informed consent was waived by the research ethics board for this study given that it was a medical record review and all patient-identifying data were made anonymous. No reporting guidelines were followed.

Inclusion criteria included all patients with primary OPSCC who presented at the OCC over the study duration who had sufficient clinical, radiology, or pathology reports in EHRs to provide TNM staging. All imaging was redone if not performed at this center and reviewed in house by qualified head and neck radiologists. There were no age-excluding criteria. Exclusion criteria included patients with OPSCC outside the indicated study duration, those that lacked sufficient reports for staging purposes, and any recurrences or patients who underwent repeat radiation therapy.

Alongside annotating the dataset to include text evidence for TNM staging from EHRs, a rule book consisting of 135 rules was established across the 4 factors that determine overall OPSCC stage: 17 rules for p16 status, 60 rules for T stage, 40 rules for N stage, and 18 rules for M stage. Rules consisted of additional guidelines that took note of synonymous terminology for critical staging structures, such as genioglossus, hyoglossus, styloglossus, and palatoglossus, which all refer to extrinsic tongue muscles. Furthermore, guidelines were also provided to sort ambiguous language that suggested uncertainty found in reports. In addition, guidelines were provided to classify stage where there was a lack of information related to size and involvement of upstaging structures but some description of primary location and extent.

To turn these rules into AI models, a supervised approach to establish a one-to-one relationship between each rule and a sentence would typically be used. However, due to rule book complexity and the limited amount of data in relation to the number of rules, a novel approach was implemented. The approach involved using the main general concepts of the rule book: size, location, state (ie, extent of involvement or extension onto or into structures), metastases, and p16 status. Training was performed using these concepts, and 4 models were

Figure. Flowchart for Machine Learning Algorithm



The flowchart describes the machine learning algorithm used to generate tumor (T), nodal (N), and metastases (M) stage for patients with oropharyngeal cancer.

[a]A numerical extraction model to determine size.

[b]An entity relationship extraction model to determine the anatomical location of the lesion and invasion state.

[c]A sequential classification model to determine if the tumor metastasized.

created, which were later reconstructed back to the rules. Training for size, location, state, and metastases was performed using existing data. To ensure that the classified stage would be that at diagnosis, only documents dated prior to the start of treatment were considered by models. For p16 status, study data were required to be used for training and validation due to a lack of existing data with documented p16 status. To do this, 500 patients selected at random from our cohort were used for training the p16 status model and 303 patients were kept unseen for validation.

The first model, referred to as model T, was an entity relationship extraction model used to assess the anatomical location of the lesion and its invasion state. The second model, known as model S, was a numerical extraction model and was trained to extract information about the size of the lesion. The third model, named model M, was a sequential classification model trained to determine if the tumor had metastasized. The p16 model filtered all sentences mentioning p16 status and aggregated them at the patient level. Methods used to generate staging models for patients with OPSCC are described in the **Figure**.

T stage was determined using model T with model S upstaging when required due to the size of the lesion. N stage was also determined using model T with model S upstaging when required, but only sentences identified as nodal were used for classification. The M staging process involved using model M to determine whether the lesion was metastatic and then using model T to provide the specific anatomical location, which was then used to determine whether the metastasis could be classified as distant. We determined p16 status using only the p16 model. It is important to note that patients with unknown p16 status (due to the lack of its documentation) were staged according to p16-negative guidelines in the ground truth and thus labeled p16 negative by the algorithm.

Overall stage was calculated based on TNM classifications and using the p16 status of the ground truth. For purposes of this study, localized cancer consisted of stage 1 and 2 cancers, while advanced cancer consisted of stage 3 and 4 cancers.

All staging results obtained were validated against the ground truth, and overall accuracy was reported as the primary evaluation metric. Precision and recall were also reported for binary features: p16 status and M stage. Precision is calculated as the number of true-positive predictions divided by the total number of positive predictions. Recall is calculated as the number of true-positive predictions divided by the sum of true positives and false negatives. Data were analyzed using R Studio version 4.2.1 (R Project for Statistical Computing) from March to November 2023.

Results

Among 806 patients with oropharyngeal cancer (mean [SD] age at the time of new disease, 63.6 [10.6] years; 651 males [80.8%]), 421 patient cancers (52.2%) were associated with HPV as indicated by p16 positivity. The distribution of clinical TNM and overall stages provided by expert reviewers as the ground truth were further outlined. Patient characteristics and demographic information for the 806 patients with oropharyngeal cancer presented at the OCC are in **Table 1**.

Table 2 presents the accuracy of the algorithm in classifying TNM stage, determining p16 status, and calculating overall stage compared with the ground truth. T stage was validated in 801 patients and correctly classified by the algorithm in 430 patients (53.7%; 95% CI, 50.2%-57.1%). When the same stage number was considered a match without specification of letter subtypes (ie, any T4 without specification of T4a or T4b), 448 patients (55.9%; 95% CI, 52.5%-59.3%) were correctly classified to the level of the stage number. Overall, N stage was validated in 806 patients and correctly classified by the algorithm in 412 patients (51.1%; 95% CI, 47.7%-54.6%). N stage of p16-positive OPSCC was validated in 421 patients and correctly classified in 284 patients (67.5%; 95% CI, 62.8%-71.8%), a greater staging accuracy than for p16-negative OPSCC, which was validated in 385 patients and correctly classified in 128 patients (33.3%; 95% CI, 28.7%-38.1%). The accuracy of N stage overall and in patients who were p16 negative increased when the same stage number was considered a match without specification of letter subtype (ie, any N2, without specification of N2a, N2b, or N2c or any N3, without specification of N3a or N3b). For overall N stage, 451 of 806 validated patients (56.0%; 95% CI, 52.5%-59.4%) were correctly classified to the level of the stage num-

Table 1. Patient Characteristics

Characteristic	Patients with OPSCC, No. (%) (N = 806)
Age at new disease, mean (SD), y	63.6 (10.6)
Sex	
Female	155 (19.2)
Male	651 (80.8)
p16 Status[a]	
Negative or unknown	385 (47.8)
Positive	421 (52.2)
Clinical T stage[a,b]	
1	175 (21.7)
2	246 (30.5)
3	106 (13.2)
4	111 (13.8)
4a	131 (16.3)
4b	32 (4.0)
Missing	5 (0.6)
Clinical N stage[a,b]	
0	106 (13.2)
1	302 (37.5)
2	112 (13.9)
2a	8 (1.0)
2b	82 (10.2)
2c	98 (12.2)
3	19 (2.4)
3a	8 (1.0)
3b	71 (8.8)
Clinical M stage[a]	
0	769 (95.4)
1	37 (4.6)
Clinical stage[a]	
1	212 (26.3)
2	122 (15.1)
3	148 (18.4)
4	319 (39.6)
Missing	5 (0.6)

Abbreviations: M, metastases; N, nodal; OPSCC, oropharyngeal squamous cell carcinoma; T, tumor.

[a] Based on manually curated ground truth.

[b] Stages T4, N2, and N3 apply to patients who were p16 positive, while stages with letter groups (T4a/b, N2a/b/c, and N3a/b) apply to patients who were p16 negative in accordance with the American Joint Committee on Cancer eighth edition cancer staging manual.

ber. For p16-negative N stage, 167 of 385 validated patients (43.4%; 95% CI, 38.5%-48.4%) were correctly classified to the level of the stage number.

M stage was validated in 806 patients and correctly classified in 706 patients (87.6%; 95% CI, 85.1%-89.7%), while p16 status was validated in 303 patients and correctly classified in 279 patients (92.1%; 95% CI, 88.5%-94.6%). Overall stage was validated in 801 patients and correctly classified in 502 patients (62.7%; 95% CI, 59.3%-66.0%). Differentiation between localized cancer stages (1 and 2) and advanced cancer stages (3 and 4) was validated in 801 patients and correctly classified in 646 patients (80.7%; 95% CI, 77.8%-83.2%).

Table 3 summarizes precision and recall for binary outputs of M stage and p16 status of the algorithm compared with the ground truth. Precision and recall for M stage were 23.1% (95% C, 18.5%-28.3%) and 73.0% (95% CI, 55.9%-86.2%), respectively. Precision and recall were higher for p16 status, with values of 91.6% (95% CI, 86.6%-94.8%) and 92.8% (95% CI, 87.5%-96.4%), respectively.

Discussion

In this diagnostic study, an AI engine had fair to good accuracy in classification of T and N stages, at 53.7% and 51.1%, respectively. Accuracy of these stage groups increased to 55.9% and 56.0% for T and N stage, respectively, when the stage number of the algorithm matched that of the ground truth without specification of letter subtype. The accuracy of N stage classification in HPV-associated disease was twice that of non-HPV-associated disease, at 67.5% and 33.3%, respectively. Accuracy increased to 43.4% in non-HPV-associated disease when the same stage number was matched without specification of letter subtype. Increases in accuracy observed when stage numbers matched suggest that the higher accuracy of p16-positive N stages may be attributed to the AJCC 8th edition guidelines, which state that p16-positive staging includes stage numbers N0, N1, N2, and N3 and p16-negative staging includes subtypes N2a, N2b, N2c, N3a, and N3b in addition to N0 and N1.[8] Although the accuracy of the AI engine was lower in T and N staging classification, the engine had greater success in classification of M stage and p16 status, with accuracy rates of 87.6% and 92.1%, respectively. Furthermore, the accuracy in calculating overall cancer stage by combining ground truth p16 status and algorithm TNM classifications was 62.7%; however, the algorithm had great success in distinguishing between localized and advanced stage cancers, with an accuracy of 80.7%.

Previously, AI has been trained to provide TNM cancer staging in individuals with lung cancers using the text of pathology reports. Nguyen et al[2] achieved accuracy at 72%, 78%, and 94% for T, N, and M staging, respectively, while McCowan et al[1] reported accuracy of 74% and 87% for T and N staging, respectively. Results from our study show comparable M staging accuracy but lower T and N staging accuracy (Table 2). Several factors may have been associated with the success rates of T and N staging, including additional complex staging guidelines from expert reviewers and clarity of radiology and clinical reports. Clinical and radiology reports were heavily used in T and N staging, while pathology reports were mostly reserved for p16 status due to a lack of sufficient pathology reports for staging. However, our experience demonstrated that radiology reports often did not comment specifically on some key factors required to radiographically stage a patient's cancer.

Errors from voice-recognition systems used in dictations contribute to staging errors because upstaging structures could be easily mistaken for similar-sounding structures. Examples include "intrinsic" vs "extrinsic" and "prevertebral" vs "paravertebral" vs "perivertebral." Occasionally, language suggesting uncertainty was observed in clinical and radiology re-

Table 2. Accuracy of Artificial Intelligence–Determined Stage vs Ground Truth

Outcome	Correct to stage number and letter level		Correct to stage number level	
	Patients matched, No./validation set	Accuracy, % (95% CI)	Patients matched, No./validation set	Accuracy, % (95% CI)
T stage	430/801	53.7 (50.2-57.1)	448/801	55.9 (52.5-59.3)
N stage				
All	412/806	51.1 (47.7-54.6)	451/806	56.0 (52.5-59.4)
p16+	284/421	67.5 (62.8-71.8)	284/421	67.5 (62.8-71.8)
p16−	128/385	33.3 (28.7-38.1)	167/385	43.4 (38.5-48.4)
M stage	706/806	87.6 (85.1-89.7)	706/806	87.6 (85.1-89.7)
p16 stage	279/303	92.1 (88.5-94.6)	279/303	92.1 (88.5-94.6)
Stage 1 vs 2 vs 3 vs 4[a]	502/801	62.7 (59.3-66.0)	502/801	62.7 (59.3-66.0)
Stage 1 and 2 vs 3 and 4[a]	646/801	80.7 (77.8-83.2)	646/801	80.7 (77.8-83.2)

Abbreviations: M, metastases; N, nodal; T, tumor.

[a] To validate stages for all patients, including those used in the training of p16 status, the overall stage was calculated using the T, N, and M stage determined by the artificial intelligence algorithm in conjunction with the ground truth p16 status.

ports, which could potentially lead to upstaging of patient cancers. When reviewers were concerned with voice-recognition errors or language with uncertainty, imaging present in patient files was reviewed to confirm whether a patient's cancer should be upstaged. Unfortunately, the algorithm is limited in its ability to make such decisions and confirm them through image analysis; however, such limitations are important to highlight as future algorithms are developed given that these algorithms could potentially incorporate both text and image analysis.

With respect to nodal status, one challenge was determining extranodal extension (ENE). Clinical reports of fixed nodes or those tethered to adjacent structures were sometimes incongruent between clinicians and imaging features and so were sometimes considered incorrect interpretation of ENE by expert reviewers. Often, large and conglomerate nodes that are cystic may feel like ENE on physical examination but prove not to be on imaging. Thus, ENE was strictly confirmed with radiological findings except for skin invasion, which was considered ENE based on clinical notes. It should be noted that even radiology reports are not the most accurate resource for determining pathologic ENE; however, very few of our patients had pathology reports to help guide staging. Furthermore, it is difficult to use radiology reports in our models because a single form of imaging is not sufficient on its own. For example, it was observed that magnetic resonance imaging (MRI) scans overcalled tumor involvement for T4b structures. Computed tomography scans were more reliable for tumor involvement, except for small structures (ie, epiglottis and vallecula), which had to be clinically correlated or could result in inaccurate upstaging of tumors. Furthermore, positron emission tomography (PET) and MRI scans were documented in additional staging guidelines to be best for determining nodal involvement and PET scans best for determining metastases. Detailed hierarchies were outlined in additional staging guidelines provided by expert reviewers stating which reports should be used in order of highest to lowest priority depending on the stage (TNM) being assessed, alongside notes of synonymous terminology for staging-related structures and phrases calling for caution when staging; these guidelines brought great strength to the algorithm.

The higher accuracy of M stage and p16 status may come from the binary nature of their classification. Given

Table 3. Precision and Recall of M Stage and p16 Status vs Ground Truth

Outcome	% (95% CI)[a]	
	Precision	Recall
M stage	23.1 (18.5-28.3)	73.0 (55.9-86.2)
p16 Status	91.6 (86.6-94.8)	92.8 (87.5-96.4)

Abbreviation: M, metastases.

[a] Precision and recall are provided only for features that are binary, such as p16-positive vs p16-negative status and M. Precision is calculated as the number of true-positive predictions divided by the total number of positive predictions. Recall is calculated as the number of true-positive predictions divided by the sum of the true positives and false negatives.

that metastatic disease defines cancer that has spread, it often does not go missing in patient workup. It is either present or absent and receives comments accordingly. Similarly, the presence of the p16 biomarker in pathology workup in HPV testing clearly notes its presence or absence. Additionally, p16 status became a requirement in the modified eighth edition of the AJCC cancer staging guidelines,[8,9] thus explaining its higher accuracy as it became mandatory to report. With these binary outputs, there is less uncertainty in categorization compared with T and N staging, which are more complex in staging guidelines and require more text data input and integration from clinical, radiology, and pathology reports.

Ultimately, study findings suggest that an AI algorithm can be successful at differentiating between localized and advanced cancer stages and thus may be used as an adjunct to clinical documentation. The clinical significance to high-accuracy differentiation of localized vs advanced stages is that disease is often diagnosed at advanced stages, where prognosis and survival outcomes are poorer compared with diagnosis at localized stages.[10,11] One study[11] reported associations between head and neck cancer stages and their 6-month survival, which indicated that individuals with more advanced cancer stages (eg, stage 4) had increased likelihood of death by nearly 4-fold compared with those with localized stages (eg, stage 1). These results suggest that localized stages of cancer have improved treatment outcomes and that high-accuracy differentiation may be able to improve on concerns noted by Crosby et al[10] surrounding improper or overtreatment, which is associated with significant morbidity.

Future steps for the algorithm may include model refinement to improve accuracy of T and N staging and expansion to other cancer centers to obtain a larger dataset for training. It will be important to test this algorithm in other health care contexts, including different electronic health records, different practices (ie, private vs academic), and regions with different health care structures (ie, regionalized vs nonregionalized cancer care). The importance of automating TNM staging goes beyond improving stage documentation rates given that accuracy of staging is important to the design of treatment regimens and toxic effects exposure. As mentioned previously, higher-accuracy staging provided by cancer registries is unavailable at the start of treatment and exhausts human resources.[3] Time allocated to human abstraction in this study was approximately 6 months for analysis of all reports, providing stage classifications, and reviewing patients with cancer as needed in regular meetings. This time may be reduced using AI. Developing AI models took considerable time up front, but all subsequent AI processing can be completed in minutes for each new patient. The algorithm used in this study can provide a preliminary stage in patients for whom stage is missing. This is certainly better than no stage at all. This can then be refined or checked by the team, particularly for variables (T and N) that are more nuanced. Future work can help create synoptic reporting by radiology to improve the accuracy of T and N staging based on those reports.

Currently, clinicians use their best judgment to determine the best treatment for a particular patient. Staging is only one part of that process, and appropriate documentation of staging is thus critical for more than just a treatment decision. It is helpful for discussion of prognosis with patients, may alter routine follow-up schedules and frequency of posttreatment imaging, and most importantly is a very basic health quality metric in an oncology practice. In all, to provide optimal treatment to patients with oropharyngeal cancer, complete staging must be documented before or at approximately the start of treatment, which AI technology has the potential to do efficiently and accurately in a timely manner. The true value of AI in the management of OPSCC is in the potential to improve patient experience by allowing clinicians to focus on the human aspect of care and less so on the technical aspect of care by reducing workload in areas like staging.

Limitations

This study has several limitations. It must be interpreted within the context of the study design. Lack of a larger dataset for training of additional rules provided by expert reviewers may explain the lower T and N staging accuracy, which in turn is associated with overall cancer stage. Furthermore, individuals in this study with unknown HPV status were staged according to HPV-negative guidelines for purposes of accounting for poor prognosis compared with their counterparts who were HPV positive. Given the increased incidence of HPV-positive disease in OPSCC,[9] many patients with unknown HPV status may be HPV positive, which can introduce bias. Nonetheless, this would not change T or M stage significantly, and while it may impact N stage, p16-positive N staging had greater accuracy provided by the algorithm and was thus easier to stage. Therefore, by staging patients with unknown HPV status according to HPV-negative guidelines, a conservative estimate for the accuracy of the algorithm for staging these cancers was provided. Future studies should exclude individuals with unknown HPV status to improve internal validity and homogeneity of data.

Conclusions

In this diagnostic study, natural language processing techniques were used to develop a machine learning algorithm capable of extracting p16 data, determining TNM stages, and calculating overall cancer stage from clinical, radiology, and pathology reports. The accuracy of T and N stage classification by the algorithm was low due to an insufficient dataset in the training of additional rules; however, binary outputs of M stage and p16 status achieved high accuracy.

Differentiation between localized and advanced stages of cancer was successful, and findings suggest this may be used clinically for documentation. With further refinement and training of models with larger datasets and improved radiographic synoptic reporting, the accuracy and applicability of the algorithm clinically may aid in timely staging before the start of treatment and may be useful in bringing the attention of clinicians to inconsistencies or disagreements in staging.

ARTICLE INFORMATION

Accepted for Publication: April 2, 2024.

Published Online: May 16, 2024. doi:10.1001/jamaoto.2024.1201

Author Affiliations: Department of Otolaryngology–Head and Neck Surgery, Sunnybrook Health Sciences Centre, University of Toronto, Toronto, Ontario, Canada (Baran, Lee, Enepekides, Higgins, Eskander); Pentavere Research Group Inc, Toronto, Ontario, Canada (Aviv, Weiss, Pettengell); Department of Radiation Oncology, Odette Cancer Centre, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada (Karam, Bayley, Poon); Division of Medical Oncology, Odette Cancer Centre, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada (Chan, Parmar, Smoragiewicz); Department of Pathology, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada (Klieb, Truong); Department of Medical Imaging, Sunnybrook Health Sciences Centre, Toronto, Ontario, Canada (Maralani).

Author Contributions: Dr Eskander had full access to all of the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.
Concept and design: Baran, Lee, Pettengell, Karam, Poon, Klieb, Higgins, Eskander.
Acquisition, analysis, or interpretation of data: Baran, Lee, Aviv, Weiss, Pettengell, Karam, Bayley, Poon, Chan, Parmar, Smoragiewicz, Truong, Maralani, Enepekides, Higgins, Eskander.
Drafting of the manuscript: Baran, Lee, Karam, Bayley, Poon, Higgins, Eskander.
Critical review of the manuscript for important intellectual content: Baran, Lee, Aviv, Weiss, Pettengell, Karam, Bayley, Poon, Chan, Parmar, Smoragiewicz, Klieb, Truong, Maralani, Enepekides, Eskander.
Statistical analysis: Weiss, Pettengell, Poon.
Obtained funding: Higgins.
Administrative, technical, or material support: Baran, Lee, Karam, Poon, Higgins, Eskander.
Supervision: Lee, Karam, Smoragiewicz, Higgins, Eskander.

Conflict of Interest Disclosures: Dr Karam reported receiving personal fees from the EMD Serono advisory board outside the submitted work. No other disclosures were reported.

Meeting Presentation: This study was presented at the AHNS Annual Meeting at COSM 2024; May 16, 2024; Chicago, Illinois.

Data Sharing Statement: See Supplement 2.

REFERENCES

1. McCowan IA, Moore DC, Nguyen AN, et al. Collection of cancer stage data by classifying free-text medical reports. *J Am Med Inform Assoc*. 2007;14(6):736-745. doi:10.1197/jamia.M2130

2. Nguyen AN, Lawley MJ, Hansen DP, et al. Symbolic rule-based classification of lung cancer stages from free-text pathology reports. *J Am Med Inform Assoc*. 2010;17(4):440-445. doi:10.1136/jamia.2010.003707

3. Evans TL, Gabriel PE, Shulman LN. Cancer staging in electronic health records: strategies to improve documentation of these critical data. *J Oncol Pract*. 2016;12(2):137-139. doi:10.1200/JOP.2015.007310

4. Vidula N, Peppercorn J. Clicking away to capture cancer staging—the benefits and challenges of completing standardized staging modules. *JCO Oncol Pract*. 2023;19(10):835-838. doi:10.1200/OP.23.00500

5. Kim BJ, Merchant M, Zheng C, et al. A natural language processing program effectively extracts key pathologic findings from radical prostatectomy reports. *J Endourol*. 2014;28(12):1474-1478. doi:10.1089/end.2014.0221

6. Cognetti DM, Weber RS, Lai SY. Head and neck cancer: an evolving treatment paradigm. *Cancer*. 2008;113(7)(suppl):1911-1932. doi:10.1002/cncr.23654

7. Eskander A, Husain ZA. HPV testing: a quality metric. *Oral Oncol*. 2020;103:104549. doi:10.1016/j.oraloncology.2019.104549

8. Amin MB, Edge SB, Greene FL, et al, eds. *AJCC Cancer Staging Manual*. 8th ed. Springer; 2017.

9. Lechner M, Liu J, Masterson L, Fenton TR. HPV-associated oropharyngeal cancer: epidemiology, molecular biology and clinical management. *Nat Rev Clin Oncol*. 2022;19(5):306-327. doi:10.1038/s41571-022-00603-7

10. Crosby D, Bhatia S, Brindle KM, et al. Early detection of cancer. *Science*. 2022;375(6586):eaay9040. doi:10.1126/science.aay9040

11. Talani C, Mäkitie A, Beran M, Holmberg E, Laurell G, Farnebo L. Early mortality after diagnosis of cancer of the head and neck—a population-based nationwide study. *PLoS One*. 2019;14(10):e0223154. doi:10.1371/journal.pone.0223154