American Head and Neck Society - Journal Club
Volume 58, March 2026

Hosted by the AHNS Artificial Intelligence Task Force (Prepared by: Antoine Eskander, MD Sc,M., FRCSC – University of Toronto; Janice L. Farlow, MD PhD – Indiana University; Rusha Patel, MD – University of Oklahoma

Table of Contents – ***click the page number to go to the summary and full article link.***

Development and Validation of Machine Learning Models for Predicting Occult Nodal Metastasis in Early-Stage Oral Cavity Squamous Cell Carcinoma

Farrokhian N, Holcomb AJ, Dimon E, Karadaghy O, Ward C, Whiteford E, Tolan C, Hanly EK, Buchakjian MR, Harding B, Dooley L, Shinn J, Wood CB, Rohde SL, Khaja S, Parikh A, Bulbul MG, Penn J, Goodwin S, Bur AM

JAMA Network Open, April 1, 2022

ABSTRACT

Background: Current standard, based on randomized clinical trial evidence, suggests that patients would benefit from an elective neck dissection (END) for oral cavity cancer (cN0), unless they are early stage and have tumors with thin depth of invasion (DOI) <4 mm. However, the vast majority of patients that have an END are pathologically N0 (pN0). As such there is added morbidity in a large proportion of patients who receive END. This study sought to develop and validate predictive models of occult nodal metastasis and to compare their performance to the DOI cutoff alone.

Methods: Seven tertiary care academic medical centers across the US provided data on adult patients with early stage (cT1-T2) from 2000-2019. Data from 6 centers was used to develop the model and was randomly split into 80% as the training set and 20% as the internal validation set. The 7th centers data (Nebraska) was then used to externally validate the cohort. Models evaluated included a standard logistic regression, random forest, support vector machine classifier and the XGBoost. XGBoost is a popular boosting algorithm particularly when machine learning is being used with small sample sizes. Different modeling approaches were compared; XGBoost, tumor depth threshold model, standard logistic regression, support vector machine classifier and random forest models.

Results: The final cohort at 634 patients. Occult nodal metastases were detected in 18% (n=114) of patients the vast majority of which were captured at time of END (n=94) with the remaining having a regional recurrence with neck observation within 2 years of their oral cavity surgery (n=20). Occult nodal metastases were associated with LVI, PNI, margin positivity, poorly differentiated tumors and greater DOI. The XGBoost supervised machine learning model outperformed all models (ROC 0.84) including

the standard logistic regression (ROC 0.78). The XGBoost model has a negative predictive value of 97.8% and a sensitivity of 91.7%.

Conclusions: This diagnostic predictive modeling study found that a machine learning algorithm (XGBoost) outperformed a standard logistic model and that both outperformed the currently used clinical threshold (DOI > 4mm). This has the potential to decrease the number of unnecessary neck dissections in patients with early-stage oral cavity cancer.

Summary
This is a wonderful study that demonstrates the power of collaboration amongst tertiary academic head and neck centers. It also highlights the value of well curated large multicenter datasets. With this highly granular data, machine learning approaches can be compared to common simplified clinical decision rules (DOI 4 mm), standard statistical approaches (logistic regression) and machine learning algorithms (such as XGBoost). The statistical approaches used were extremely thoughtful and performed at a high level. Ultimately the XGBoost Machine Learning approach demonstrates the best diagnostic accuracy. With this approach, the number needed to screen to correctly identify additional patients with pN positive disease is 21 and the number needed to scree to correctly identify additional patients with pN negative disease (avoiding END) would be 11.

Strengths
Relatively large sample size study that is multicenter with both an internal validation and an external validation approach. The statistical methods are sound and robust. The comparative approach of the simplistic DOI cutoff which is often used clinically to standard (logistic regression) statistical approaches and to the more novel machine learning (XGBoost) model is thoughtful and appropriate.

Weaknesses
Due to the years of the study (2000-2019) many patients did not have data on depth of invasion and in those patients, tumor thickness was used instead of DOI. More importantly, there is already a randomized trial study looking at this question. The authors very correctly identify that the mechanism for survival benefit in that randomized study may be related to occult disease that was resected but because of a typical bisection pathology assessment (as opposed to a breadloaf assessment as would be done with a sentinel node), patients labelled as pN0 may have had micrometastatic disease that was not identified on pathology. As such, it is unlikely that a retrospective study with or without a machine learning approach would change the conclusions/results of that randomized trial. Also, within that clinical trial, many subgroups were identified as higher risk for occult nodal disease (buccal and floor of mouth subsite, poorly differentiated tumor) and as such if the machine learning algorithm was compared to a more complex clinical decision tool that accounted for these additional clinical factors, it is unlikely that the predictive abilities would be much improved. This is partially demonstrated by the fact that the standard logistic regression model (ROC 0.78) performed similarly to the XGBoost algorithm (0.84).

back to top

Screening for extranodal extension in HPV-associated oropharyngeal carcinoma: evaluation of a CT-based deep learning algorithm in patient data from a multicentre, randomised de-escalation trial

Benjamin H Kann, Jirapat Likitlersuang, Dennis Bontempi, Zezhong Ye, Sanjay Aneja, Richard Bakst, Hillary R Kelly, Amy F Juliano, Sam Payabvash, Jeffrey P Guenette, Ravindra Uppaluri, Danielle N Margalit, Jonathan D Schoenfeld, Roy B Tishler, Robert Haddad, Hugo J W L Aerts, Joaquin J Garcia, Yael Flamand, Rathan M Subramaniam, Barbara A Burtness, Robert L Ferris

Lancet Digital Health, June 2023

ABSTRACT

Background: Pretreatment identification of pathological extranodal extension (ENE) would guide therapy de-escalation strategies for in human papillomavirus (HPV)-associated oropharyngeal carcinoma but is diagnostically challenging. ECOG-ACRIN Cancer Research Group E3311 was a multicentre trial wherein patients with HPV-associated oropharyngeal carcinoma were treated surgically and assigned to a pathological risk-based adjuvant strategy of observation, radiation, or concurrent chemoradiation. Despite protocol exclusion of patients with overt radiographic ENE, more than 30% had pathological ENE and required postoperative chemoradiation. We aimed to evaluate a CT-based deep learning algorithm for prediction of ENE in E3311, a diagnostically challenging cohort wherein algorithm use would be impactful in guiding decision-making.

Methods: For this retrospective evaluation of deep learning algorithm performance, we obtained pretreatment CTs and corresponding surgical pathology reports from the multicentre, randomised de-escalation trial E3311. All enrolled patients on E3311 required pretreatment and diagnostic head and neck imaging; patients with radiographically overt ENE were excluded per study protocol. The lymph node with largest short-axis diameter and up to two additional nodes were segmented on each scan and annotated for ENE per pathology reports. Deep learning algorithm performance for ENE prediction was compared with four board-certified head and neck radiologists. The primary endpoint was the area under the curve (AUC) of the receiver operating characteristic.

Results: From 178 collected scans, 313 nodes were annotated: 71 (23%) with ENE in general, 39 (13%) with ENE larger than 1 mm ENE. The deep learning algorithm AUC for ENE classification was 0·86 (95% CI 0·82–0·90), outperforming all readers ($p<0·0001$ for each). Among radiologists, there was high variability in specificity (43–86%) and sensitivity (45–96%) with poor inter-reader agreement (κ 0·32). Matching the algorithm specificity to that of the reader with highest AUC (R2, false positive rate 22%) yielded improved sensitivity to 75% (+ 13%). Setting the algorithm false positive rate to 30% yielded 90% sensitivity. The algorithm showed improved performance compared with radiologists for ENE larger than 1 mm ($p<0·0001$) and in nodes with short-axis diameter 1 cm or larger.

Conclusion: The deep learning algorithm outperformed experts in predicting pathological ENE on a challenging cohort of patients with HPV-associated oropharyngeal carcinoma from a randomised clinical trial. Deep learning algorithms should be evaluated prospectively as a treatment selection tool.

Summary

The study group retrospectively evaluated a CT-based deep learning algorithm (DualNet) for detecting pretreatment extranodal extension (ENE) in patients with HPV-associated oropharyngeal cancer. The study compared the algorithm's performance to that of four experienced head and neck radiologists in classifying ENE using pretreatment CT scans with segmented lymph nodes. A key finding of the study is that the deep learning algorithm significantly outperformed the expert radiologists in predicting ENE, including ENE larger than 1 mm and in nodes with a short-axis diameter of 1 cm or more. Human readers demonstrated significant inter-reader variability, and both the readers and the AI model had more difficulty with larger lymph nodes. The authors conclude that these results suggest that deep learning algorithms have the potential to be valuable screening tools for ENE in this patient population, which could aid in treatment decision-making, including the selection of appropriate de-escalation strategies and minimizing the use of trimodality therapy

Strengths:

- The data set was pulled from the ECOG 3311 trial, which required centralized pathology overread. This limited the potential for variations in pathologic classification of ENE.

- The methods for measurement (DualNet and human readers) were well selected and trained. The deep learning algorithm, DualNet was trained and retrained prior to use. The selected human readers were experienced fellowship-trained head and neck radiologists who were given a set of consistent guidelines to use when evaluating for ENE.
- The study was adequately powered. Power analysis suggested the need for 155 lymph nodes, and 313 lymph nodes were analyzed.

Weaknesses

- By nature of using ECOG 3311 data, this study excluded patients who were felt to have obvious ENE, somewhat limiting the generalizability of this data to that of the known ENE+ patient cohort.
- DualNet relied on manual segmentation of lymph nodes for analysis, which adds to the manual labor time. The alternative, whole image analysis, was not performed. This could limit the ability to implement this type of AI due to lack of experienced manpower.
- While the human readers had document experience in head and neck radiology, their comfort level with ENE may have been variable. An attempt was made to correct for this by providing consistent imaging findings for ENE, but the levels of experience between the readers may have been different and may have led to more false positive or negative results among less experienced readers.
- As designed, this trial examines a subset of oropharyngeal cancer patients. This data will be more robust when tested prospectively on in an 'all comers' population.

back to top

The Ethics of Using Artificial Intelligence in Scientific Research: New Guidance Needed for a New Tool

Resnik DB, Hosseini M

AI Ethics, April 2025

ABSTRACT
Using artificial intelligence (AI) in research offers many important benefits for science and society but also creates novel and complex ethical issues. While these ethical issues do not necessitate changing established ethical norms of science, they require the scientific community to develop new guidance for the appropriate use of AI. In this article, we briefly intro- duce AI and explain how it can be used in research, examine some of the ethical issues raised when using it, and offer nine recommendations for responsible use, including:

(1) Researchers are responsible for identifying, describing, reducing, and controlling AI-related biases and random errors;

(2) Researchers should disclose, describe, and explain their use of AI in research, including its limitations, in language that can be understood by non-experts;

(3) Researchers should engage with impacted communities, populations, and other stakeholders concerning the use of AI in research to obtain their advice and assistance and address their interests and concerns, such as issues related to bias;

(4) Researchers who use synthetic data should
(a) indicate which parts of the data are synthetic;
(b) clearly label the synthetic data;
(c) describe how the data were generated; and
(d) explain how and why the data were used;

(5) AI systems should not be named as authors, inventors, or copyright holders but their contributions to research should be disclosed and described;

(6) Education and mentoring in responsible conduct of research should include discussion of ethical use of AI.

COMMENTARY
The bioethicists Drs. Resnik and Hosseini provide a comprehensive review of the ethical implications of using AI in scientific research, including the discussion of benefits and challenges, needed guidelines, and recommendations for responsible AI use. They focus on "narrow AI," which is designed for specific tasks. Ethical issues raised include systemic and random errors, which can undermine the quality and trustworthiness of research. AI also has a "black box" problem, where its lack of transparency impacts understandability and accountability. Data privacy, including the potential disclosure of data input into AI, is important both for intellectual property as well as research participant privacy. Finally, moral agency, or whether and how AI contributions should be acknowledged as authors and investigators, has also been called into question. Based on these ethical issues, the authors propose recommendations that could enhance the accountability, objectivity, reproducibility, rigor, transparency, social responsibility, and fairness of AI use in scientific research.

back to top

Large Language Model Influence on Diagnostic Reasoning: A Randomized Clinical Trial

Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P J Olson, Adam Rodman, Jonathan H Chen

JAMA Network Open, October 2024

ABSTRACT

Importance: Large language models (LLMs) have shown promise in their performance on both multiple-choice and open-ended medical reasoning examinations, but it remains unknown whether the use of such tools improves physician diagnostic reasoning.

Objective: To assess the effect of an LLM on physicians' diagnostic reasoning compared with conventional resources.

Design, Setting & Participants: A single-blind randomized clinical trial was conducted from November 29 to December 29, 2023. Using remote video conferencing and in-person participation across multiple academic medical institutions, physicians with training in family medicine, internal medicine, or emergency medicine were recruited.

Intervention: Participants were randomized to either access the LLM in addition to conventional diagnostic resources or conventional resources only, stratified by career stage. Participants were allocated 60 minutes to review up to 6 clinical vignettes.

Main Outcomes & Measures: The primary outcome was performance on a standardized rubric of diagnostic performance based on differential diagnosis accuracy, appropriateness of supporting and opposing factors, and next diagnostic evaluation steps, validated and graded via blinded expert consensus. Secondary outcomes included time spent per case (in seconds) and final diagnosis accuracy. All analyses followed the intention-to-treat principle. A secondary exploratory analysis evaluated the standalone performance of the LLM by comparing the primary outcomes between the LLM alone group and the conventional resource group.

Results: Fifty physicians (26 attendings, 24 residents; median years in practice, 3 [IQR, 2-8]) participated virtually as well as at 1 in-person site. The median diagnostic reasoning score per case was 76% (IQR, 66%-87%) for the LLM group and 74% (IQR, 63%-84%) for the conventional resources-only group, with an adjusted difference of 2 percentage points (95% CI, −4 to 8 percentage points; P = .60). The median time spent per case for the LLM group was 519 (IQR, 371-668) seconds, compared with 565 (IQR, 456-788) seconds for the conventional resources group, with a time difference of −82 (95% CI, −195 to 31; P = .20) seconds. The LLM alone scored 16 percentage points (95% CI, 2-30 percentage points; P = .03) higher than the conventional resources group.

Conclusions & Relevance In this trial, the availability of an LLM to physicians as a diagnostic aid did not significantly improve clinical reasoning compared with conventional resources. The LLM alone demonstrated higher performance than both physician groups, indicating the need for technology and workforce development to realize the potential of physician-artificial intelligence collaboration in clinical practice.

Summary

The authors present a trial of 50 physicians (multicenter; internal, family, or emergency medicine attendings and residents) randomized to using ChatGPT 4.0 vs conventional diagnostic resources for diagnosis of 6 cases adapted from a landmark study used to evaluate computer-based diagnostic systems. Participants filled out a structured reflection grid to provide rationale for a list of differential diagnoses and next steps. ChatGPT was separately tested on its diagnostic capabilities, and blinded raters scored both human and ChatGPT responses. They found that diagnostic accuracy and time spent for the task did not differ between the randomized groups, but ChatGPT alone performed better than either human group. The authors conclude that untrained LLM use in this simulated environment did not improve physician diagnostic performance, but as LLMs alone outperformed both physician groups, further development in human-computer interactions is needed for AI-assisted diagnostic reasoning.

Strengths

- The randomized trial design with blinded expert rating of a validated structured reflection tool, applied to tested clinical cases not published on the internet (i.e. and thus not available for LLM training data) adds rigor and reliability

- The breakdown of diagnostic reasoning via structured reflection and a recording of time provides a nuanced analysis of diagnostic reasoning

Weaknesses

- Although this was a multicenter design of both attendings and residents from several "general" medical fields, the small sample size and number of clinical cases limits generalizability
- Participants were not trained in prompt engineering or other specifics of LLM use, which mimics real world scenarios but limits some of the potential advantages of LLM use
- The simulation lacks contextual details of the actual clinical environment that may impact clinical decision making
- Only one LLM was tested at a specific point in time, which limits generalizability to other potential LLMs or AI technology built into clinical decision support systems

back to top