Article:
Abou-Foul, A.K., Kristunas, C., Henson, C. et al. (2026) Correlation between head and neck radiologists reporting on extranodal extension detected on radiological imaging: a head and neck cancer international group multinational study. Oral Oncology, 178. 107980. ISSN: 1368-8375

https://doi.org/10.1016/j.oraloncology.2026.107980

Oral Oncology 178 (2026) 107980

Contents lists available at ScienceDirect

Oral Oncology

journal homepage: www.elsevier.com/locate/oraloncology

Correlation between head and neck radiologists reporting on extranodal extension detected on radiological imaging: A head and neck cancer international group multinational study

A.K. Abou-Foul [a,1], C. Kristunas [b,1], C. Henson [c], D. Andrew [d], F. Bidault [e], J.W. Dankbaar [f], E.S. Gebrim [g], P. de Graaf [h], J. Kim [i], A.D. King [j], H. Kuno [k], S. Medrano-Martorell [l], J.P.N. Goh [m], E. Santos-Armentia [n], D.G. Schafigh [o], X.C. Wu [p], Y. Xiao [q], S.H. Huang [r], W.M. Lydiatt [s], E. Yu [t,1], B.F. Branstetter [u,2], H. Mehanna [a,2,*]

[a] *Institute of Head and Neck Studies and Education, University of Birmingham, B15 2TT, UK*
[b] *Institute for Head and neck studies and education, University of Birmingham, Birmingham, UK*
[c] *Department of Radiation Oncology, Stephenson Cancer Center, University of Oklahoma Health Sciences Centre, Oklahoma City, OK, USA*
[d] *School of Clinical Dentistry, University of Sheffield, Sheffield, UK*
[e] *Department of Medical Imaging, Gustave Roussy Cancer Campus, Paris-Saclay University, Villejuif, France, and Multimodal Biomedical Imaging Laboratory (BioMaps), Paris-Saclay University, Paris, France*
[f] *Radiology department, University Medical Center, Utrecht, Netherlands*
[g] *Department of Radiology, Faculdade de Medicina da Universidade de Sao Paulo, Brazil*
[h] *Department of Radiology and Nuclear Medicine, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands, and Cancer Center Amsterdam, Imaging and Biomarkers, Amsterdam, the Netherlands*
[i] *Radiology department, University of Hong Kong, Hong Kong Special Administrative Region, China*
[j] *The Chinese University of Hong Kong, Department of Imaging and Interventional Radiology Prince of Wales Hospital, Hong Kong S.A.R., China*
[k] *Department of Diagnostic Radiology, National Cancer Center Hospital East, Kashiwa, Japan*
[l] *Department of Radiology, Hospital Clinic de Barcelona, Spain*
[m] *Department of Diagnostic Radiology, Tan Tock Seng Hospital, Singapore*
[n] *Radiology Department, Ribera Povisa Hospital, Vigo, Spain*
[o] *Department of Otorhinolaryngology, Head and Neck Surgery, University of Cologne, Germany*
[p] *Department of Radiology and Biomedical Imaging, University of California San Francisco, San Francisco, CA, USA*
[q] *Department of Radiology, Clinical Oncology School of Fujian Medical University & Fujian Cancer Hospital (Fujian Branch of Fudan University Affiliated Cancer Hospital), Fuzhou, Fujian Province, China*
[r] *Department of Radiation Oncology, Princess Margaret Cancer Centre/University of Toronto, Canada*
[s] *Department of Surgery, Creighton University School of Medicine, Omaha, Nebraska, USA, and Methodist Estabrook Cancer Center, Nebraska Methodist Hospital, Omaha, Nebraska, USA*
[t] *Department of Neuroradiology and Head and Neck Imaging, Princess Margaret Cancer Centre/University of Toronto, Canada*
[u] *Department of Radiology, University of Pittsburgh, Pittsburgh, PA, USA*

ARTICLE INFO

Keywords:
Extra nodal extension
Extracapsular spread

ABSTRACT

Introduction: Extranodal extension detected on imaging (iENE) is an increasingly recognised risk factor for poor prognosis in patients with head and neck squamous cell carcinoma. However, its clinical usefulness depends on

* Corresponding author at: Director, Institute for Head and neck studies and education, University of Birmingham, Birmingham, UK.
E-mail addresses: a.abou-foul@bham.ac.uk (A.K. Abou-Foul), c.a.kristunas@bham.ac.uk (C. Kristunas), Christina-Henson@ouhsc.edu (C. Henson), Andrew@sheffield.ac.uk (D. Andrew), Francois.BIDAULT@gustaveroussy.fr (F. Bidault), J.W.Dankbaar@umcutrecht.nl (J.W. Dankbaar), eloisa.gebrim@hc.fm.usp.br (E.S. Gebrim), p.degraaf@amsterdamumc.nl (P. de Graaf), janejkim76@gmail.com (J. Kim), King2015@cuhk.edu.hk (A.D. King), hkuno@east.ncc.go.jp (H. Kuno), medrano@clinic.cat (S. Medrano-Martorell), autolycus432@gmail.com (J.P.N. Goh), esantosar@gmail.com (E. Santos-Armentia), darius.schafigh@uk-koeln.de (D.G. Schafigh), xin.wu@ucsf.edu (X.C. Wu), xyp999@yeah.net (Y. Xiao), ShaoHui.Huang@uhn.ca (S.H. Huang), William.Lydiatt@nmhs.org (W.M. Lydiatt), Eugene.Yu@uhn.ca (E. Yu), bfb1@pitt.edu (B.F. Branstetter), h.mehanna@bham.ac.uk (H. Mehanna).
[1] Equal first authorship.
[2] Equal senior authorship.

https://doi.org/10.1016/j.oraloncology.2026.107980
Received 12 January 2026; Received in revised form 6 April 2026; Accepted 18 April 2026
Available online 26 April 2026

Radiology
Head and neck cancer
Squamous cell carcinoma
Classification system
Inter-rater correlation

accurate and reliable radiological assessment. A major limitation to the use of iENE as a prognostic marker remains variability in its interpretation, attributable in part to the lack of standardised diagnostic criteria.
Methods: This international, multicentre reliability study included expert head and neck radiologists nominated to represent member groups of the Head and Neck Cancer International Group (HNCIG). Participants underwent standardised training using the HNCIG consensus classification and educational atlas, and then independently evaluated anonymised CT and MRI scans using both binary (iENE present or absent) and ordinal (grades 0–3) classifications over two rounds. Inter- and intra-rater reliability were evaluated using free-marginal multi-rater kappa statistics and intraclass correlation coefficients. Subgroup analyses were performed by imaging modality and radiologist experience.
Results: Fourteen radiologists representing 12 research groups, spanning 25 countries, completed both assessment rounds. Overall inter-rater agreement was moderate for binary classification ($\kappa = 0.75$, 95% CI 0.66–0.84) and weak for ordinal classification ($\kappa = 0.49$, 95% CI 0.42–0.56). Intra-rater agreement was moderate for binary classification and moderate to strong for ordinal classification. Agreement levels did not differ significantly by imaging modality or years of experience.
Conclusion: Reliable radiological assessment is necessary for the use of iENE in clinical trials, staging, and routine decision making. Using the HNCIG consensus classification in a binary format produced consistent results across international expert radiologists and imaging modalities, supporting its wider use. However, substantial variability remains in grading the extent of iENE, indicating the need for further training support.

Introduction

Extranodal extension (ENE), defined as the extension of metastatic tumour beyond the lymph node capsule, is a well-established adverse prognostic factor in locally advanced head and neck squamous cell carcinoma (HNSCC). Its presence on post-operative histopathological examination is associated with increased risk of locoregional recurrence and distant metastasis, warranting treatment intensification with adjuvant chemoradiotherapy.[1].

Emerging evidence suggests that ENE identified on pre-treatment imaging (imaging-detected ENE, iENE) is also prognostically relevant in both HPV-independent and HPV-associated oropharyngeal squamous cell carcinoma (OPSCC).[2–5] Consequently, iENE has been included in the 9th edition of the TNM stage classifications (AJCC version 9 and UICC 9th edition).[6,7] Clinically, iENE may influence therapeutic strategy, for example, guiding the selection of primary chemoradiotherapy over surgery to avoid the morbidity of trimodal therapy. Conversely, if validated as a robust prognostic marker, iENE could identify patients suitable for intensification strategies, such as neoadjuvant chemo-immunotherapy.[8].

The utility of iENE hinges on its validity and interobserver reliability. A recent multinational study demonstrated wide inter-institutional variability in iENE accuracy rates in the real-world setting, limiting its prognostic reliability.[9] Computed Tomography (CT) scanning generally confers greater sensitivity, whereas magnetic resonance imaging (MRI) provides higher specificity. However, combining both modalities improved diagnostic accuracy considerably.[9] Moreover, diagnostic accuracy correlated with the radiologist's experience in head and neck cancer imaging. Meta-analyses have similarly shown inconsistent diagnostic performance for iENE on CT and MRI (pooled sensitivity 60–77% specificity 60–96%), with widely varying inter-observer agreement (correlation coefficients 0.4–0.7).[4,5].

One key contributor to variability is the absence of standardised diagnostic criteria. A review by Huang *et al.* highlighted the heterogeneity in iENE definitions across studies.[10] In response, the Head and Neck Cancer International Group (HNCIG, https://www.hncig.org)—a global consortium of 23 national clinical trials organisations—undertook a Delphi process to develop a consensus four-grade iENE classification system, now endorsed by 19 professional groups across 39 countries.[11].

This study aimed to assess the inter- and intra-observer agreement among specialist radiologists from international centres who had been trained in the interpretation of iENE on cross-sectional imaging modalities when applying the HNCIG consensus classification for the diagnosis of iENE.

Methods

Study design and participants

A steering group formed under the auspices of HNCIG, comprising experts in head and neck radiology, surgery, oncology, and biostatistics, designed and coordinated the study (supplementary Table A). Each of the 23 HNCIG member groups was invited to nominate up to two expert radiologists. Nominees were required to be experienced radiologists, who were recognised national or international experts in head and neck imaging, and to commit to completing both study rounds. Twelve groups participated, including organisations from Europe, Asia, and the Americas (supplementary Table B). All participants were included as manuscript co-authors.

Training and scoring Protocols

Participants were provided with and asked to study the HNCIG consensus classification and the accompanying educational atlas[11], which included 46 annotated examples across all grades (0–3) of iENE. They then completed a self-test set of ten CT and MRI scans, and then reviewed the provided results, expert interpretations, and scoring rationale. This served as training and calibration prior to scoring the formal study set.

In the first round, participants scored a blinded study set comprising a mix of anonymised 10 CT and 30 MRI scans, using both binary (iENE present/absent) and ordinal (grade 0–3) HNCIG classifications. Relevant images in multiple planes (or reformats) were provided to the observers. Only images that included neck lymph nodes were included, to maintain efficiency for the observers. Six weeks later, a second round of reviews was conducted with a total of 41 CT and MRI scans, including a re-presentation of 5 CT and 13 MRI scans from round one and 23 new CT scans, to enable intra-rater reliability analysis. Each round was open for 28 days, with two reminder emails sent. Participants were blinded to their own previous scores and to others' scores. Two experienced radiologists from the steering committee reviewed together the scans with low interrater agreement to elucidate possible reasons for the inconsistency between raters.

As this was a retrospective study using anonymised, routinely acquired clinical images, formal ethical approval was not required according to the lead institution's research ethics committee (University Hospitals Birmingham NHS Foundation Trust).

Outcomes

The primary endpoint was inter-rater agreement across all 63 scans for binary and grade classification, quantified using the free-marginal multi-rater kappa statistic[12]. Secondary endpoints included intra-rater agreement across both scoring rounds for each radiologist quantified using the free-marginal multi-rate kappa statistic[12], and inter- and intra-rater agreement quantified by the intraclass correlation coefficient (ICC).

Statistical analysis

Intra- and inter-rater agreement was assessed using kappa statistics, which measures the agreement beyond what would be expected by chance. Due to a high prevalence of "present" cases in the binary classification and since there were no restrictions on raters' distribution of cases to categories, the free-marginal multi-rater kappa statistic[12] was selected, with 95% confidence intervals estimated using Gwet's variance formula[13] and implemented using Randolph's online calculator[14]. As a secondary measure of inter- and intra-rater agreement, intra-class correlation coefficients (ICCs) were also estimated based on a single-rater, absolute-agreement, two-way model[15] and implemented using the *icc* command from the *irr* package in *R*. The free-marginal multi-rater kappa was preferred over the ICC for primary analyses, as the ICC considers categories to be ordered linearly, which cannot be assumed for the grade classification. A crude measure of agreement, the percentage agreement is also presented for the inter-rater analysis.

Intra-rater agreement was assessed for each participant separately, then averaged (mean and median). Pooled estimates were calculated, by combining classifications across all participants.

Subgroup analyses were performed by imaging modality (CT vs MRI) and years of radiologist experience (<20 year, ≥20 years).

For interpretation of results, we followed the following conventions: A kappa value of < 0.6 indicates weak to no agreement, whereas a kappa of 0.6–0.79, 0.8–0.9 or > 0.9 indicate moderate, strong or almost perfect agreement respectively.[16] For ICC, values of < 0.5, 0.5–0.75, 0.75–0.9 or > 0.9 indicate poor, moderate, good or excellent agreement respectively.[17].

Assuming an alpha of 0.05, and proportional categories, 40 cases would provide over 80% power to detect a difference in kappa of 0.4 (0.3 vs 0.7) for either classification, and 90% power to detect a difference in kappa of 0.3 (0.5 vs 0.8) for the grade classification.[18] However, the proportionality assumption was found to be violated for the binary classification after the first round, and so additional cases were added to balance the proportions of "present" and "absent" cases, in addition to providing more CT cases for the purpose of subgroup analyses. With 63 cases, we expect to have over 80% power to detect a difference in kappa of 0.3 (0.5 vs 0.8) for either classification, and over 90% power to detect a difference in kappa of 0.2 (0.7 vs 0.9) for the grade classification.[18].

Results

Fourteen expert head and neck radiologists representing 12 research groups, spanning 25 countries, participated in this study. The full list is provided in supplementary Table B. In total, 12/14 (85.7%) participants completed both rounds and 11/12 provided the number of years since they graduated from medical school. One participant withdrew after the first round and was not substituted. One participant completed only the second round. The participating radiologists had between 9 and 41 years of experience in head and neck radiology; 36% (n = 4) had less than 20 years of experience, the remaining 64% (n = 7) had 20 or more years of experience.

Inter-rater agreement

The free-marginal multi-rater kappa between all 12 radiologists for all 63 scans using the HNCIG criteria was moderate, 0.75 (95%CI: 0.66, 0.84) for binary classification (iENE positive/negative), and weak, 0.49 (95%CI: 0.42, 0.57), for Grade classification (0–3), see Fig. 1 and supplementary Table C. The classification proportions for each of the 63 scans are presented in Fig. 2: 15.9% (10/63) scans showed a difference of 20% or more in the proportion of radiologists scoring iENE positive versus negative. ICC for the overall study set was moderate, 0.51 (95% CI: 0.42, 0.61), for binary classification and also for Grade, 0.65 (95%CI: 0.56, 0.74). The percentage agreement for the whole set was 87.45% (95%CI: 86.40, 88.42) for binary classification and 61.95% (95%CI: 60.47, 63.42) for Grade.

Sensitivity analyses were performed by imaging modality and radiologists' experience. When considering the 33 CT scans and 30 MRI scans separately, analyses for both modalities demonstrated a similar level of agreement to the overall set, with CT scans generally having slightly better agreement than MRI (supplementary Table C). For binary classification, free marginal kappa was 0.77 (95%CI: 0.65, 0.89) and 0.73 (95%CI: 0.59, 0.86) respectively for CT and MRI scans. ICC was 0.52 (95%CI: 0.39, 0.66) and 0.51 (95%CI: 0.38, 0.66) respectively and percentage agreement was 88.43% (95%CI: 87.02, 89.71) and 86.36% (95%CI: 84.78, 87.80) for CT and MRI respectively. Similarly for grade classification, free marginal kappa was 0.50 (95%CI: 0.39, 0.61) and 0.48 (95%CI: 0.39, 0.58) respectively for CT and MRI scans. ICC was 0.63 (95%CI: 0.51, 0.75) and 0.67 (95%CI: 0.55, 0.79) respectively and percentage agreement was 62.58% (95%CI: 60.53, 64.59) and 61.26% (95%CI: 59.10, 63.38) for CT and MRI respectively.

Separating the group by number of years since graduating medical school showed similar agreement among the two groups: those with ≥ 20 (n = 7) years of experience and those with < 20 (n = 4) years of experience. However there appeared to be a greater degree of agreement among those < 20 years of experience. For binary classification, the < 20 years group had a free-marginal kappa of 0.84 (95%CI: 0.74, 0.93), ICC of 0.62 (95%CI: 0.51, 0.73) and percentage agreement of 91.80% (95%CI: 88.59, 94.16). For the ≥ 20 years group these measures were 0.74 (95%CI: 0.65, 0.83), 0.49 (95%CI: 0.39, 0.60) and 87.00% (95%CI: 85.08, 88.70) respectively. For the grade classification, the < 20 years group had a free-marginal kappa of 0.58 (95%CI: 0.48, 0.68), ICC of 0.75 (95%CI: 0.66, 0.83) and percentage agreement of 68.52% (95% CI:63.67, 72.99). For the ≥ 20 years group these measures were 0.48 (95%CI: 0.40, 0.56), 0.62 (95%CI: 0.52, 0.72) and 61.30% (95%CI: 58.65, 63.89) respectively (Appendix Table C).

Intra-rater agreement

The 12 radiologists each rated twice a total of 18 (5 CT and 13 MRI) studies. When examining intra-rater correlation for the binary classification (see Table 1 and Fig. 3), the free-marginal kappa ranged from weak to perfect agreement (0.44 to 1.00) across the 12 radiologists, with on average moderate agreement [mean 0.76 (95%CI: 0.67, 0.85), median 0.78 (IQR: 0.67, 0.89)]. When pooled across all radiologists the free-marginal kappa was moderate, 0.76 (95%CI: 0.67, 0.85). For grade classification, the marginal kappa suggested moderate agreement: range was 0.33–0.93, mean was 0.63 (95%CI: 0.54, 0.73), median was 0.63 (IQR: 0.56, 0.71) and the pooled kappa was 0.63 (95%CI: 0.55, 0.71). Fig. 3 shows that free-marginal multi-rater kappa using the binary classification was higher than when using the grade classification for 10 of the 12 raters, and when the ratings were pooled. However, there was still wide variation in intra-rater agreement per rater for both binary and grade classifications.

ICCs for the binary classification ranged between poor and excellent agreement (0.274–1.00) and on average were moderate (mean 0.67 (95%CI: 0.55, 0.79), median 0.65 (IQR: 0.55, 0.85)). The pooled ICC showed moderate agreement, 0.65 (95%CI: 0.57, 0.72). For the grade classification, the ICC suggested good agreement: range was: 0.51–0.98, mean was at 0.78 (95%CI: 0.70, 0.87), median was 0.77 (IQR: 0.65, 0.92) and the pooled ICC was good, 0.77 (95%CI: 0.71, 0.83).

Subgroup analysis by scan type demonstrated that the level of agreement in MRI scans alone did not differ from overall (CT and MRI

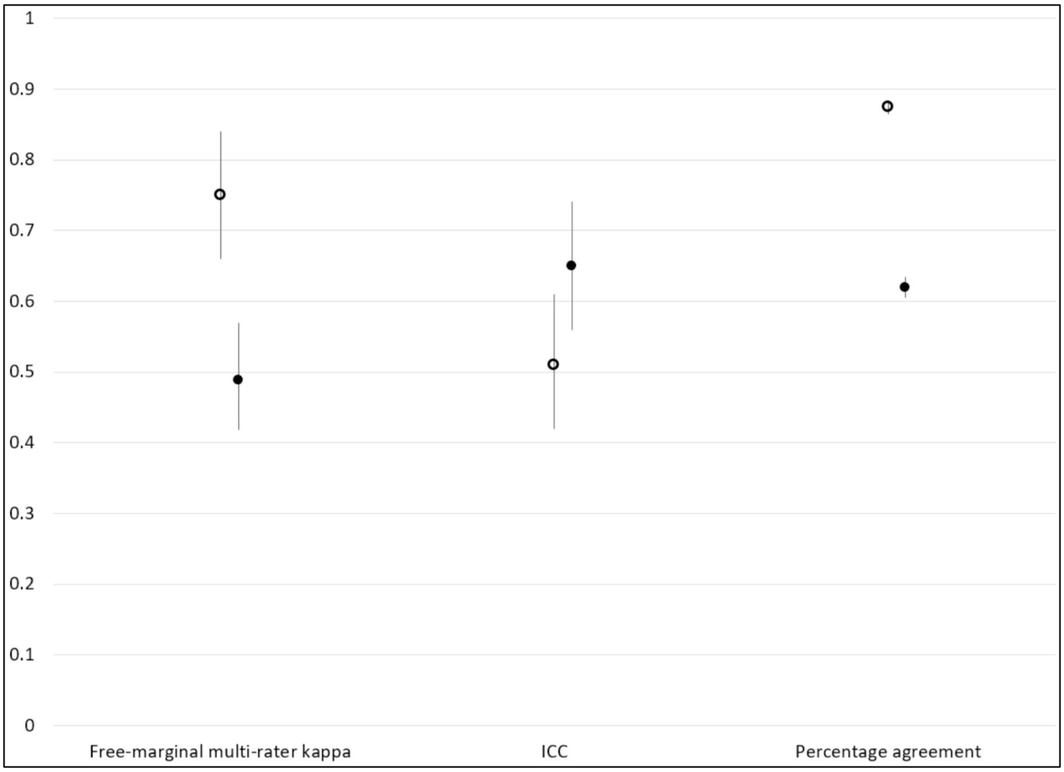

Fig. 1. Inter-rater agreement within the 12 radiologists, measured by free-marginal multi-rater kappa, Intraclass correlation coefficient (ICC) and percentage agreement, for binary (hollow circle) and grade (solid circle) classifications. Bars show 95% confidence intervals).

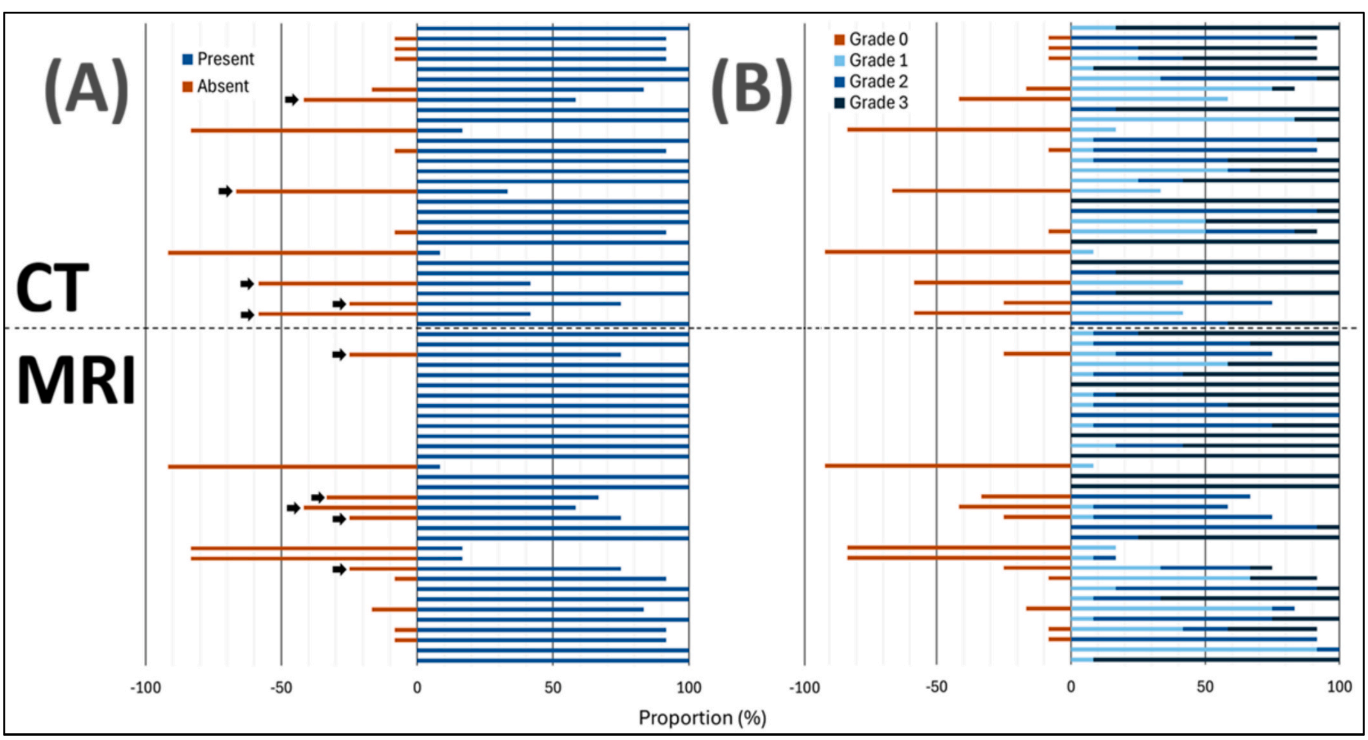

Fig. 2. Classification proportion of radiologists for each scan (CT top, MRI bottom) by (A) binary and (B) grade classifications. Scans resulting in ≥ 20% variation between iENE positive and iENE negative are indicated by an arrow.

scans together) (Appendix Table D). Intra-rater analyses for MRI alone (13 scans) for binary classification by free marginal kappa showed moderate agreement: range was 0.38–1.00, mean was 0.71 (95%CI: 0.58, 0.83), median was 0.69 (IQR: 0.54, 0.85) and pooled kappa was 0.71 (95%CI: 0.59, 0.82). For Grade, free-marginal kappa also showed moderate agreement: range was 0.28–0.90, mean was 0.64 (95%CI: 0.53, 0.75), median was 0.64 (IQR: 0.54, 0.79) and pooled kappa was 0.64 (95%CI: 0.55, 0.73). ICCs for the binary classification by MRI alone

Table 1
Intra-rater agreement of 12 radiologists each rating 18 images (5 CT and 13 MRI) twice using a binary (present/absent) or grade (0–3) classification. (*<20 years' experience,^experience data unavailable).

Rater	Free-marginal kappa		ICC	
	Binary (present/absent)	Grade (0–3)	Binary (present/absent)	Grade (0–3)
1^	0.67 (0.31, 1.00)	0.56 (0.26, 0.85)	0.67 (0.31, 0.86)	0.65 (0.27, 0.85)
2*	1.00 (1.00, 1.00)	0.85 (0.65, 1.00)	1.00 (1.00, 1.00)	0.96 (0.88, 0.98)
3	0.67 (0.31, 1.00)	0.56 (0.26, 0.85)	0.59 (0.20, 0.82)	0.63 (0.26, 0.84)
4	0.89 (0.67, 1.00)	0.63 (0.35, 0.91)	0.88 (0.71, 0.95)	0.74 (0.42, 0.89)
5	0.78 (0.48, 1.00)	0.93 (0.78, 1.00)	0.62 (0.25, 0.84)	0.98 (0.94, 0.99)
6*	0.44 (0.02, 0.87)	0.63 (0.35, 0.91)	0.27 (−0.23, 0.65)	0.76 (0.45, 0.90)
7*	0.89 (0.67, 1.00)	0.63 (0.36, 0.91)	0.83 (0.61, 0.93)	0.90 (0.76, 0.96)
8	0.89 (0.67, 1.00)	0.63 (0.35, 0.91)	0.86 (0.67, 0.95)	0.79 (0.52, 0.92)
9	0.67 (0.31, 1.00)	0.33 (0.02, 0.65)	0.51 (0.09, 0.78)	0.65 (0.28, 0.85)
10	0.89 (0.67, 1.00)	0.78 (0.54, 1.00)	0.78 (0.51, 0.91)	0.93 (0.81, 0.97)
11*	0.78 (0.48, 1.00)	0.63 (0.35, 0.91)	0.61 (0.21, 0.84)	0.89 (0.73, 0.96)
12	0.56 (0.16, 0.95)	0.41 (0.09, 0.72)	0.41 (−0.01, 0.72)	0.51 (0.08, 0.78)
Mean (95% CI):	**0.76 (0.67, 0.85)**	**0.63 (0.54, 0.73)**	**0.67 (0.55, 0.79)**	**0.78 (0.70, 0.87)**
Median (IQR):	**0.78 (0.67, 0.89)**	**0.63 (0.56, 0.71)**	**0.65 (0.55, 0.85)**	**0.77 (0.65, 0.92)**
Range:	**0.44–1.00**	**0.33–0.93**	**0.27–1.00**	**0.51–0.98**
Pooled	**0.76 (0.67, 0.85)**	**0.63 (0.55, 0.71)**	**0.65 (0.57, 0.72)**	**0.77 (0.71, 0.83)**
<20 years' experience*				
Mean (95% CI):	**0.75 (0.64, 0.87)**	**0.69 (0.59, 0.78)**	**0.68 (0.41, 0.94)**	**0.88 (0.81, 0.95)**
Median (IQR):	**0.78 (0.67, 0.84)**	**0.63 (0.63, 0.74)**	**0.72 (0.44, 0.92)**	**0.90 (0.83, 0.93)**
Range:	**0.56–0.89**	**0.63–0.85**	**0.27–1.00**	**0.76–0.96**
Pooled:	**0.78(0.63, 0.92)**	**0.69 (0.55, 0.82)**	**0.63 (0.46, 0.75)**	**0.87 (0.80, 0.92)**
≥20 years' experience				
Mean (95% CI):	**0.77 (0.64, 0.89)**	**0.61 (0.47, 0.75)**	**0.66 (0.54, 0.79)**	**0.75 (0.63, 0.86)**
Median (IQR):	**0.78 (0.67, 0.89)**	**0.63 (0.41, 0.78)**	**0.62 (0.51, 0.86)**	**0.74 (0.63, 0.93)**
Range:	**0.44–1.00**	**0.33–0.93**	**0.41–0.88**	**0.51–0.98**
Pooled:	**0.76 (0.65, 0.88)**	**0.61 (0.50, 0.71)**	**0.57 (0.39, 0.70)**	**0.75 (0.61, 0.84)**

showed moderate agreement: range was 0.29–1.00, mean was 0.66 (95%CI: 0.53, 0.79), median was 0.63 (IQR: 0.51, 0.83) and pooled ICC was 0.65 (95%CI: 0.55, 0.73). ICCs for grade showed good agreement: range was 0.46–0.97, mean was 0.79 (95%CI: 0.69, 0.89), median was 0.86 (IQR: 0.66, 0.93) and the pooled ICC was 0.78 (95%CI: 0.71, 0.83). Intra-rater analyses for CT scan were unreliable due to small numbers.

Subgroup analysis by experience of radiologist demonstrated slightly better agreement in those with < 20 years' experience than those with ≥ 20 years' experience, Table 1. Intra-rater analyses for < 20 years (n = 4) for binary and grade classification by pooled free-marginal kappa showed moderate agreement: 0.78 (95%CI: 0.63, 0.92) and 0.69 (95% CI: 0.55, 0.82) respectively. For those with ≥ 20 years' experience (n = 7) for binary and grade classification, intra-rater analysis by pooled free-marginal kappa also showed moderate agreement: 0.76 (95%CI: 0.65, 0.88) and 0.61 (95%CI: 0.50, 0.71) respectively.

Qualitative analysis of scans with low interrater agreement

In light of the results, our two steering committee radiologists (BFB, EY) reviewed nine scans that showed lowest inter-rater agreement to elucidate possible reasons for the inconsistency between raters. The main themes were defining matted nodes (five scans), distinguishing displacement from invasion of muscle (two scans), whether 270-degree encirclement of the carotid artery is required to qualify as invasion (one scan), and not recognising skin as an organ (one scan).

Discussion

This international study demonstrates that the binary HNCIG consensus classification for iENE enables moderate inter- and intra-rater agreement among expert head and neck radiologists for both CT and MRI. These agreement levels were at the higher end of those reported by prior studies, which were usually single-centre studies that would be expected to show higher correlations.[19] Classification of grade showed weak inter-rater agreement and weak to moderate intra-rater agreement. Surprisingly, even when assigning binary grades (iENE positive vs negative), 15% of scans where graded differently by 20% or more of radiologists. Furthermore, whilst overall there was moderate to high intra-rater agreement, radiologists demonstrated wide intra-rater variability for individual scans. These findings demonstrate the reliability of the HNCIG iENE classification when used for binary diagnosis, and support its suitability for clinical research and potential clinical implementation. However, these also highlight the subjective nature of iENE grading in some cases.

Notably, inter-rater agreement was slightly higher for CT than MRI, but differences were not statistically significant. This aligns with prior literature suggesting CT provides better sensitivity and MRI better specificity, with complementary roles in practice.[5,9] Our results support the use of either modality when applying the consensus classification, particularly when reviewed by experienced radiologists. Subgroup analysis showed slightly higher agreement among radiologists with < 20 years' experience. Nevertheless, all participants achieved good agreement, suggesting that the classification can be reliably applied by radiologists with a range of experience after appropriate calibration.

Assigning specific grades showed statistically significantly lower inter-rater agreement (weak 0.49 (0.42, 0.57)) than a simple binary classification (moderate 0.75 (0.66, 0.84)). This may be expected since the larger the number of options, the more likely the variability, and potentially the higher the subjectivity in applying the grades. However, the large decrease in agreement is notable. Fortunately, clinical practice and decision making in head and neck squamous cancer currently require the identification of iENE, and do not in the main depend on iENE grade.

Even when assigning binary grades, there were some scans that caused significant variability between raters. Fig. 2 showed that 15% of scans resulted in differences of 20% or more in the proportion of radiologists assigning no iENE versus iENE positive, meaning that at least 20% of patients in those cases would have been misclassified. This has direct clinical impact on treatment decision making, as generally patients with iENE are considered ineligible for surgical treatment. Our two steering committee radiologists reviewed these scans to elucidate possible reasons for the inconsistency between raters. The main themes that arose included: what constitutes matted nodes, difference between displacement and invasion of muscle, 270-degree encirclement of carotid artery to qualify as invasion, and not recognising skin is an organ. Reasons for this variability need to be examined and researched further.

We had expected that intra-rater correlation would be very high, due to high internal consistency, and the free-marginal multi-rater kappa for the binary classification did show moderate to strong agreement. However, there was high variability with a wide range of intra-rater agreement amongst individual scans read by the same radiologist.

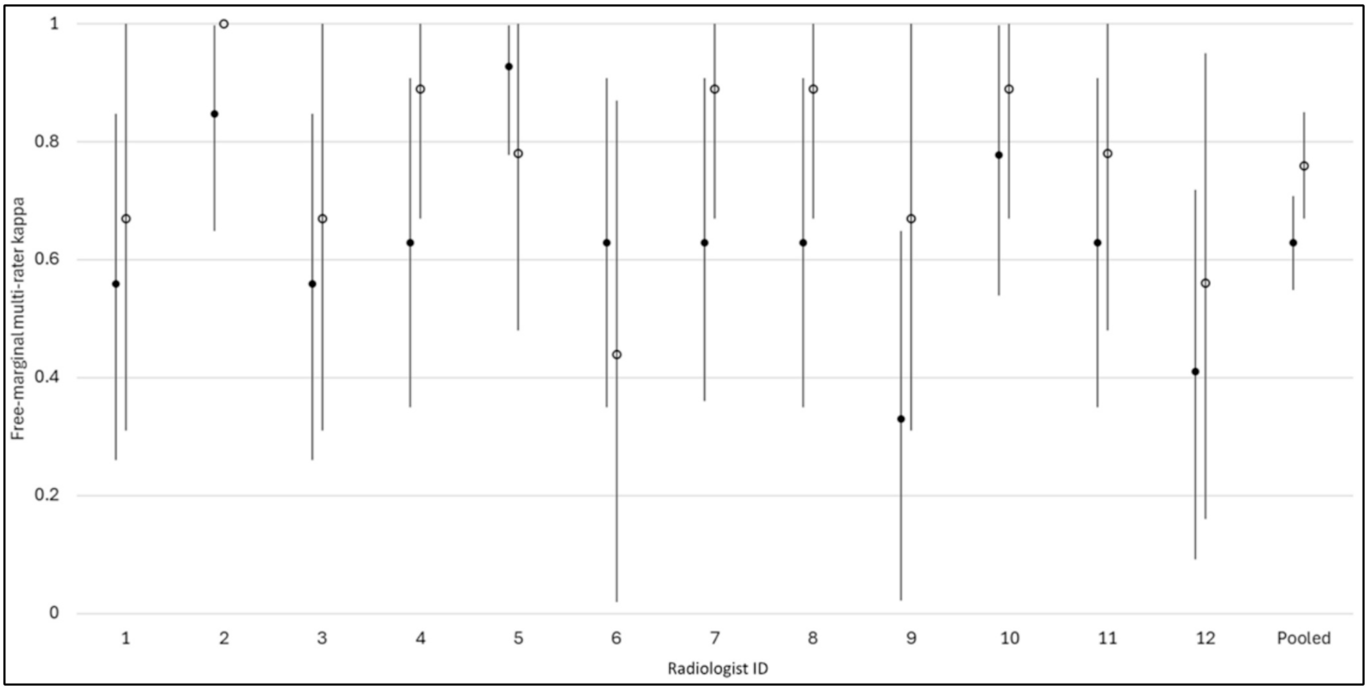

Fig. 3. Intra-rater agreement within the 12 radiologist and overall pooled free-marginal multi-rater kappa for binary (hollow circle) and grade (solid circle) classifications.

Furthermore, intra-rater agreement was weak for some radiologists, particularly for the grade classification, where a third of the radiologists showed weak agreement. These findings are surprising and highlight the subjective nature of diagnosing iENE, especially in some cases. Reasons for these inconsistencies and means for their mitigation should be studied further.

An additional consideration is the evolving treatment landscape, particularly the increasing use of neoadjuvant immunotherapy in head and neck cancer. Our findings demonstrate that iENE assessment already exhibits variability, particularly in cases with ambiguous or borderline imaging features. In this context, neoadjuvant immunotherapy may further complicate interpretation, as treatment-related changes in nodal architecture, including immune infiltration, fibrosis, and necrosis, may further complicate the radiologic assessment of ENE. This further emphasises the importance of standardised iENE criteria, and the need for future studies to understand their performance in treatment-modified nodal disease.

The clinical relevance of iENE is also context-dependent and varies across tumour subsites. In predominantly surgically managed cancers, such as in the oral cavity, pathological ENE remains the key determinant of adjuvant treatment. However, the role of iENE is distinct, as it may inform pre-treatment decision making. In disease settings where both surgical and non-surgical approaches are considered, such as in oropharyngeal cancer, suspected iENE may influence the selection of primary treatment in order to reduce the likelihood of trimodality therapy.

Tumour subsite data was not systematically recorded in this study, which limits formal subsite-specific analysis. However, the cases were derived from routine clinical imaging datasets and were likely predominantly oropharyngeal cancers, where imaging-detected ENE has the greatest clinical relevance in guiding initial treatment selection.

The strengths of this study include its international scope, rigorous methodology, and use of real-world CT and MRI scans. The two-round design allowed evaluation of intra-rater consistency. We also applied robust statistical methods, including sensitivity analyses using ICC. Finally, one of the challenges of implementing any iENE classification internationally is that it must be easily applicable with limited training, as it is not realistic or possible to provide all radiologists with specific intensive training. We did not assess the performance of the radiologists before reading and applying the HNCIG consensus classification. However, we were able to demonstrate that by studying the online HNCIG consensus classification[11], its accompanying educational Atlas (downloadable from https://www.hncig.org), and then taking the self-test, head and neck radiologists from different countries were able to provide consistent results across diverse imaging modalities.

Our study has limitations. One potential limitation could be the absence of histopathological correlation, preventing assessment of diagnostic accuracy. This was by design, as our aim was to evaluate reliability, not validity. Second, the dataset was limited in size and may not capture the full heterogeneity of imaging quality or tumour biology seen in clinical practice. However, scans were carefully selected to include a representative range of presentations, and the high-level agreement across centres strengthens generalisability.

Consensus criteria for iENE diagnosis are one important component of improving iENE reliability, accuracy and prognostic power. A recent study demonstrated other potential complementary strategies.[9] Experienced radiologists showed better specificities, with similar sensitivities to general radiologists. Our study only recruited experienced specialist head and neck radiologists. Length of experience in this group of specialist head and neck radiologists did not appear to make a difference. But it is possible that correlation may be lower between general radiologists. Another strategy was the use of both MRI and CT together, which was shown to considerably improve both sensitivity and specificity of iENE. This may also improve the consistency and correlation between radiologists, and requires further study, including its cost effectiveness.

The quantitative and qualitative work in this study suggested that additional training materials are needed to ensure optimal reproducibility of iENE assessment between radiologists. Further online resources, based on the HNCIG classification, have now been developed by the International joint Task Force on iENE[20], with the aim of providing additional training in a widely accessible manner, that should help to further enhance consistency and reliability. These training materials should be validated further by international studies representing head and neck radiologists from around the world. In the future, the application of artificial intelligence-driven radiomics algorithms may also improve the consistency and possibly accuracy of reporting iENE. [21] The HNCIG has embarked on the Artificial Intelligence in Head and Neck Cancer (ArI) initiative to collect datasets that will enable the development and validation of such artificial intelligence driven algorithms for iENE identification amongst others.

In conclusion, consistent radiologic scoring is essential for incorporating iENE into future clinical trials, staging systems and routine clinical decision making. Application of the HNCIG consensus classification (in a binary format) and its accompanying training materials yielded consistent results across diverse international expert radiologists and imaging modalities, supporting the reliability of the HNCIG iENE consensus classification for international adoption. There is however wide variation in grading extent of iENE, and more is needed to address this. To enhance consistency further, using the guidelines in conjunction with recently published online training resources is recommended.

Contributors.

HM, AKA-F, BFB, EY and CH conceived the study concept and initiated the study design. HM, BFB, EY, CH, AKA-F, were involved in study development, data analysis, and data interpretation of this Policy Review. CK performed the data analysis. All authors participated in data generation and collection, manuscript preparation, and approved the final manuscript.

All other authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

The data that support the findings of this study are available from the corresponding author, [HM], upon reasonable request.

Conflict of interest statement.

HM is a senior investigator for the National Institute for Health Research (NIHR). The views expressed in this Policy Review are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. **HM** is the director and a shareholder of Warwickshire Head and Neck Clinic; chair of the Head and Neck Cancer International Group (HNCIG); and past president of the British Association of Head and Neck Oncologists. HM also reports receiving honoraria from AstraZeneca; Speakers Bureau on MSD, Sanofi Pasteur, Merck; research Funding from GSK Biologicals, MSD, Sanofi Pasteur, GSK Plc, AstraZeneca; travel Accommodation Expenses from Sanofi, Pasteur, MSD, Merck. **WL** declares he is the chair of American Joint Committee on Cancer (AJCC) Head and Neck Core Group for the 9th Version. **SY** received research funding from Bristol-Myers Squibb and EMD Serono.CH receives payment as an ad hoc scientific advisory board member for EMD Serono.

CRediT authorship contribution statement

A.K. Abou-Foul: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **C. Kristunas:** Writing – original draft, Visualization, Software, Formal analysis. **C. Henson:** Writing – review & editing, Conceptualization. **D. Andrew:** Writing – review & editing. **F. Bidault:** Writing – review & editing. **J.W. Dankbaar:** Writing – review & editing. **E.S. Gebrim:** Writing – original draft. **P. de Graaf:** Writing – review & editing. **J. Kim:** Writing – review & editing. **A.D. King:** Writing – review & editing. **H. Kuno:** Writing – review & editing. **S. Medrano-Martorell:** Writing – review & editing. **J.P.N. Goh:** . **E. Santos-Armentia:** Writing – review & editing. **D.G. Schafigh:** Writing – review & editing. **X.C. Wu:** Writing – review & editing. **Y. Xiao:** Writing – review & editing. **S.H. Huang:** Writing – review & editing, Methodology, Conceptualization. **W.M. Lydiatt:** Writing – review & editing, Methodology, Conceptualization. **E. Yu:** Writing – review & editing, Writing – original draft, Supervision, Project administration, Methodology, Investigation, Data curation, Conceptualization. **B.F. Branstetter:** Writing – review & editing, Writing – original draft, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **H. Mehanna:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.oraloncology.2026.107980.

References

[1] Bernier J, Cooper JS, Pajak TF, van Glabbeke M, Bourhis J, Forastiere A, et al. Defining risk levels in locally advanced head and neck cancers: a comparative analysis of concurrent postoperative radiation plus chemotherapy trials of the EORTC (#22931) and RTOG (# 9501). Head Neck 2005;27(10):843–50.

[2] Benchetrit L, Torabi SJ, Givi B, Haughey B, Judson BL. Prognostic significance of Extranodal Extension in HPV-Mediated Oropharyngeal Carcinoma: a Systematic Review and Meta-analysis. Otolaryngol Head Neck Surg 2021;164(4):720–32.

[3] Lydiatt WM, Patel SG, O'Sullivan B, Brandwein MS, Ridge JA, Migliacci JC, et al. Head and Neck cancers-major changes in the American Joint Committee on cancer eighth edition cancer staging manual. CA Cancer J Clin. 2017;67(2):122-37.

[4] Morey T, Stern C, Foreman A, Hodge JC. Accuracy of imaging modalities at detecting extracapsular spread of cervical lymph node metastases in HPV-associated oropharyngeal cancer: a systematic review protocol of diagnostic test accuracy. JBI Evid Synth 2022;20(1):189–95.

[5] Park SI, Guenette JP, Suh CH, Hanna GJ, Chung SR, Baek JH, et al. The diagnostic performance of CT and MRI for detecting extranodal extension in patients with head and neck squamous cell carcinoma: a systematic review and diagnostic meta-analysis. Eur Radiol 2021;31(4):2048–61.

[6] Brierley J, Giuliani M, O'Sullivan B, Rous B, Van Eycken E. International Union Against Cancer: TNM Classification of Malignant Tumours -9th edition. 9 ed: Wiley; 2025.

[7] Evans M, Huang SH, Ho A, O'Sullivan B, Branstetter B, Lewis J, et al. AJCC Cancer Staging System: Oropharynx (HPV-Associated). 9 ed: American Joint Committee on Cancer; 2025.

[8] Sadeghi N, Mascarella MA, Khalife S, Ramanakumar AV, Richardson K, Joshi AS, et al. Neoadjuvant chemotherapy followed by surgery for HPV-associated locoregionally advanced oropharynx cancer. Head Neck 2020;42(8):2145–54.

[9] Mehanna H, Abou-Foul AK, Henson C, Kristunas C, Nankivell PC, McDowell L, et al. Accuracy and Prognosis of Extranodal Extension on Radiologic Imaging in Human Papillomavirus-Mediated Oropharyngeal Cancer: a Head and Neck Cancer International Group (HNCIG) Real-world Study. Int J Radiat Oncol Biol Phys 2025; 123(2):432–41.

[10] Huang SH, O'Sullivan B, Su J, Bartlett E, Kim J, Waldron JN, et al. Prognostic importance of radiologic extranodal extension in HPV-positive oropharyngeal carcinoma and its potential role in refining TNM-8 cN-classification. Radiother Oncol 2020;144:13–22.

[11] Henson C, Abou-Foul AK, Yu E, Glastonbury C, Huang SH, King AD, et al. Criteria for the diagnosis of extranodal extension detected on radiological imaging in head and neck cancer: Head and Neck Cancer International Group consensus recommendations. Lancet Oncol 2024;25(7):e297–307.

[12] Randolph J. Free-Marginal Multirater Kappa (multirater K[free]): An Alternative to Fleiss' Fixed-Marginal Multirater Kappa. Proceedings of the Joensuu University Learning and Instruction Symposium 2005 [Internet]. 2005. [Available from: http://files.eric.ed.gov/fulltext/ED490661.pdf]. Last accessed: 15/12/2025.

[13] Gwet KL. Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Multiple Raters. 3rd ed: Advanced Analytics, LLC; 2010.

[14] Randolph JJ. Online Kappa Calculator [Computer software] 2008. Version: Version. [Available from: http://justusrandolph.net/kappa/]. Last accessed: 15/12/2025.

[15] Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15(2):155–63.

[16] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb) 2012; 22(3):276–82.

[17] Portney LG, Watkins MP. Foundations of Clinical Research: Applications to Practice. 3rd ed: Pearson/Prentice Hall; 2009.

[18] Bujang MA, Baharum N. Guidelines of the minimum sample size requirements for Kappa agreement test. Epidemiology, Biostatistics, and Public Health. 2017;14(2): e12267.1-10.

[19] Hoebers F, Yu E, O'Sullivan B, Postma AA, Palm WM, Bartlett E, et al. Augmenting inter-rater concordance of radiologic extranodal extension in HPV-positive oropharyngeal carcinoma: a multicenter study. Head Neck 2022;44(11):2361–9.

[20] Imaging-derived Extranodal Extension (iENE) in Head and Neck Cancer: A Guide for Radiologists from AOSHNHR-ASHNR-ESHNR Joint Task Force on iENE 2025. Version: Version. [Available from: https://ashnr.org/iene/]. Last accessed: 15/12/2025.

[21] Kann BH, Likitlersuang J, Bontempi D, Ye Z, Aneja S, Bakst R, et al. 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. Lancet Digit Health 2023;5(6):e360–9.