Diagnostic accuracy of imaging modalities in detection of histopathological extranodal extension
A systematic review and meta-analysis
Abdel-Halim, Chadi Nimeh; Rosenberg, Tine; Dyrvig, Anne Kirstine; Høilund-Carlsen, Poul Flemming; Sørensen, Jens Ahm; Rohde, Max; Godballe, Christian

Published in:
Oral Oncology

DOI:
10.1016/j.oraloncology.2020.105169

Publication date:
2021

Document version:
Accepted manuscript

Document license:
CC BY-NC-ND

Citation for published version (APA):

Go to publication entry in University of Southern Denmark's Research Portal

Terms of use
This work is brought to you by the University of Southern Denmark. Unless otherwise specified it has been shared according to the terms for self-archiving. If no other license is stated, these terms apply:

• You may download this work for personal use only.
• You may not further distribute the material or use it for any profit-making activity or commercial gain
• You may freely distribute the URL identifying this open access version

If you believe that this document breaches copyright please contact us providing details and we will investigate your claim. Please direct all enquiries to puresupport@bib.sdu.dk
Abstract

Objective: To present an up to date systematic review and meta-analysis evaluating the diagnostic accuracy of the most used imaging modalities in detection of histopathological extra nodal extension (ENE) in head and neck squamous cell carcinoma.

Materials and Methods: With help of a specialized librarian, Medline, Embase, and Cochrane databases were systematically searched on March 27th 2020. Screening, inclusion, quality assessment, and data extraction were done by two reviewers. Meta-analysis was conducted using the bivariate model approach after pooling the studies according to imaging modality. Heterogeneity was explored by meta-regression. Comparison was done by meta-regression and sub-group analyses.

Results: Out of 476 initial hits, 25 studies were included for analysis. Of these, 14 dealt with CT, nine with PET/CT, four with MRI, two with ultrasound, and none with PET/MRI. Meta-analysis based on a total sample size of 3,391 showed that CT had a sensitivity of 76% [67% - 82%] and specificity of 77% [69% - 83%], MRI a sensitivity of 72% [64% - 79%] and specificity of 78% [57% - 90%], and PET/CT a sensitivity of 80% [76% - 84%] and specificity of 83% [74% - 90%] in the ability to predict ENE. No meta-analysis could be done on ultrasound. There were no significant differences between modalities in overall accuracy; however, PET/CT had significantly higher sensitivity than CT and MRI.

Conclusion: There was no significant difference in the ability of CT, MRI, and PET/CT to diagnose histopathological ENE, except that PET/CT had a significantly higher sensitivity than CT and MRI.

Keywords: Head and Neck Cancer; Oral Cancer; Larynx; Pharynx; Squamous Cell Carcinoma; Extranodal Extension; Positron-Emission tomography; X-Ray computed Tomography; Magnetic Resonance Imaging; Ultrasonography

1 Abbreviations: ENE, extranodal extension; HNSCC, head and neck squamous cell carcinoma; US, ultrasound; CT, computed tomography; MRI, magnetic resonance imaging; PET, positron emission tomography; SROC, summary receiver operating characteristic;
**Introduction**

The growth of a metastatic tumor beyond the encapsulated lymph node is referred to as extranodal extension (ENE). Detection of ENE is one of the most important prognostic factors for patients with head and neck squamous cell carcinoma (HNSCC) [1-3]. Due to its prognostic significance it was introduced in the newest edition of the TNM classification by the Union for International Cancer Control in 2017 [4]. ENE increases recurrence and mortality rates, and if detected after curatively intended surgery the treatment must be intensified, usually with chemoradiation [5]. Such triple-treatment may increase morbidity and reduce quality of life.

ENE may be clinically obvious as seen in the rare cases of cervical metastases penetrating the skin. However, in most cases, tumor growth is limited to the perinodal soft tissue and clinicians must rely on diagnostic imaging techniques. Traditionally, ultrasound (US), computed tomography (CT), magnetic resonance imaging (MRI), and positron emission tomography (PET) with CT (PET/CT) are the modalities used in the clinical setting. Several studies have examined the diagnostic accuracy among different imaging techniques [6-30]. However, the reported results vary with sensitivities ranging from 47 to 100% and specificities from 40% to 100%. No recent systematic literature reviews on diagnostic imaging of ENE in HNSCC patients currently exist. Thus, a solid evaluation is warranted.

We aimed to systematically review the literature on the diagnostic value of imaging in detection of ENE in HNSCC patients, and to perform a meta-analysis comparing accuracies among CT, MRI, US, PET/CT and PET/MRI.

**Materials and Methods**

A systematic review was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) statement [31]. The review protocol was registered in the PROSPERO International Prospective Register of Systematic Reviews with registration number CRD42020187976.
**Systematic Literature Search**

A PICO (population, intervention, comparison, and outcome) strategy was adopted for the search [32]. The target population consisted of HNSCC patients with ENE. HNSCC was defined as squamous cell carcinomas in the oropharynx, hypopharynx, larynx, oral cavity or cervical squamous cell carcinomas with unknown primary tumor. ENE was defined as histopathological extension of metastatic tumor, present within the confines of a lymph node, through the capsule into the surrounding connective tissue with or without stromal reaction [33]. Radiology in the form of PET, CT, MRI, ultrasonography, or a combination of these constituted the intervention. Histopathological evaluation served as the comparator. Outcome measures were true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) in detection of ENE on radiology compared to histology. Finally, the study design was without restrictions, as long as original data were provided.

We developed a search strategy from a combination of free text and Medical Subject Heading terms for HNSCC, lymph nodes, ENE, PET/CT, CT, PET/MRI, MRI, US, and imaging (see supplementary material 1 for the full strategy). Block search was selected as search strategy, as it accommodated the PICO approach. Terms were identified and selected with help of a scientific librarian, and on March 27th 2020 we searched the electronic bibliographic databases Embase, Medline, and Cochrane with no restriction on publication date.

**Study Selection**

Studies identified in the literature search were entered into Covidence (Veritas Health Innovation Ltd, Australia) and duplicates were removed manually. To assess the applicability of studies, two authors (CNA and TR) independently screened titles and abstracts using predefined criteria for inclusion and exclusion. The inclusion criteria in the screening process were: HNSCC, ENE, imaging, and specification of diagnostic accuracy values as TP, FP, TN, FN, sensitivity, specificity, predictive values, likelihoods ratios, odds ratios, accuracy, or ROC-analysis. Exclusion criteria were: patients without HNSCC and reviews.
Next, the same two authors independently evaluated the full text version of all studies that passed through the screening. Case reports, studies written in other languages than English, German, Norwegian, Swedish, Danish, or Turkish, and studies without relevant data (e.g. studies including non-HNSCC cancers and studies using machine learning in the evaluation of their index test) were excluded. Moreover, studies were excluded if data did not enable tabulating 2 x 2 contingency tables for the calculation of diagnostic accuracy values. At each step in the selection process, discrepancies were solved through discussion and in case of doubt, the study passed to the next step.

**Data Item Collection Process**

Data extraction was performed by two authors (CNA and TR). Data were extracted on country, study design, patient selection (age, sex, primary tumor site, HPV status, T-stage, and N-stage), imaging modality, diagnostic criteria for imaging and histopathology, analysis reference (all lymph nodes or histological malignant lymph nodes), sample unit (lymph nodes, cervical regions, necks, or patients), TP, TN, FP, and FN. In studies where diagnostic accuracies were compared between two examiners/observers, outcomes for both observers were extracted and used in analysis. In studies that reported accuracy values for different diagnostic criteria or thresholds, the outcomes giving the highest accuracy values were extracted.

**Risk of Bias Assessment**

Risk of bias and quality of individual studies were assessed using the QUADAS-2 Checklist for primary diagnostic accuracy studies structured around four domains; Patient Selection, Index Test, Reference Standard, and Flow and Timing [34]. Publications assessed by these criteria as being of unacceptable quality were excluded. Unacceptable quality was defined as studies with >2 domains with high risk of bias or for studies on PET scans >3 domains with high risk of bias (this was due to the observation that none of these studies predefined the threshold for ENE, and thus all were assessed with high risk of bias in domain 2 about the index test).
Synthesis of Results

Diagnostic accuracy in terms of sensitivity, specificity, and diagnostic odds ratio with corresponding 95% confidence intervals was calculated on the basis of TP, FP, TN, and FN extracted from each of the included studies. All data analyses were performed in Stata version 16 (StataCorp LLC, Texas, USA) in addition to the statistical software package “midas” [35]. Studies in which two observers’ outcomes were extracted, both were included in the analyses. Sensitivity was defined as the proportion of patients, necks, cervical regions, or lymph nodes correctly identified radiologically as having ENE, and specificity as the proportion of patients, necks, cervical regions, or lymph nodes correctly identified on radiology as not having ENE. Diagnostic odds ratio was defined as the odds of the imaging test being positive for a patient having ENE relative to the odds of the imaging test being positive for patients not having ENE. Accuracy was defined as the area under ROC curves (AUC).

Meta-analyses were conducted for each modality alone based on bivariate random effects mixed-models [36]. A paired forest plot was done for all studies to demonstrate study-specific sensitivity and specificity. The assumption of bivariate normality and the interdependence of sensitivity and specificity were checked graphically before analyses were done for the individual imaging modalities. The proportion of variation due to threshold effects was calculated as the squared correlation coefficient estimated from the between-study covariance parameter. Plots were made with observed data points, summary performance estimates, and summary receiver operating characteristic (SROC) curves with confidence and prediction intervals.

Heterogeneity was calculated as Higgins I², where a value of 0% indicated no observed heterogeneity and values greater than 50% were considered substantial heterogeneity. In cases where sufficient data were available, multiple univariable meta-regression was used to explore the causes of heterogeneity. Potential sources of heterogeneity tested were: publication year, study type, threshold, quality assessment, definition of histopathological ENE, sample unit (patient, neck, cervical region, or lymph...
node), and whether negative lymph nodes accounted for histopathological benign and malignant lymph nodes or only malignant nodes. To assess potential publication bias, Deek’s funnel plot for all studies was used with the slope coefficient tested against null (p-value < 0.10). Comparison of diagnostic accuracy among the imaging modalities included sub-group analyses with meta-regression on sensitivity alone, specificity alone, and both in a joint model with grouping of PET/CT vs. CT, PET/CT vs. MRI, and CT vs. MRI [36]. Statistical tests were considered significant against the null-hypothesis if p-values < 0.05.

**Results**

**Study Selection**

The search resulted in 476 studies after removal of duplicates. After screening against title and abstract, 52 studies were assessed for full-text eligibility. Twenty-five of these were included in the final review and meta-analysis. The remaining 27 studies were excluded for different reasons. Figure 1 illustrates the study selection process.

**Study Characteristics**

A total of 25 studies with 1995 patients were identified and included in the meta-analysis. In four of the studies two different imaging modalities were used. Fourteen of the studies used CT, nine used PET/CT, four used MRI, two used US, and none used PET/MRI or PET alone. Four CT-studies contained assessments from two different radiologists. Different units for study samples were used: In eight out of 25 studies, ENE detection was on patient level; in 8/25, on neck level; in 5/25, on cervical region level; and in the rest (4/25) analysis was made on the level of lymph nodes. In total, the meta-analysis included a sample size of 3391, consisting of 863 patients, 1396 necks, 891 cervical regions, and 241 lymph nodes. See Table 1 for a detailed description of the included studies.
Risk of Bias within Studies

None of the studies were of low overall quality. Seven of the studies were of high quality. High risk of bias was found in the domain of reference standard in 13 of the studies, and in the domain of the index test in seven of the studies. A summary of the results of the risk of bias assessment is presented in Table 2.

Results of Individual Studies

The diagnostic values (sensitivity and specificity) for all studies with all imaging modalities and all radiologists are presented in the forest plot in Figure 2. The plot showed substantial heterogeneity with sensitivity ranging from 47 to 100% and specificity from 40 to 100%.

Diagnostic Value of CT

The results of the meta-analysis on the diagnostic value of CT for detecting ENE showed a pooled sensitivity of 76% [67% - 82%], specificity of 77% [69% - 83%], diagnostic odds ratio of 10 [5 – 16], and AUC of 0.83 [0.79 – 0.86]. A summary ROC of the meta-analysis can be seen in Figure 3.

Diagnostic Value of MRI

MRI had a pooled sensitivity of 72% [64% - 79%], specificity of 78% [57% - 90%], diagnostic odds ratio of 9 [4 – 23], and AUC of 0.76 [0.72 – 0.80]. The summary ROC of the meta-analysis can be seen in Figure 3.

Diagnostic Value of PET/CT

PET/CT had a pooled sensitivity of 80% [76% - 84%], specificity of 83% [74% - 90%], diagnostic odds ratio of 20 [11 – 38] and AUC of 0.81 [0.78 – 0.85]. The summary ROC of the meta-analysis can be seen in Figure 3.
**Diagnostic Value of US**

Only two studies reported on the diagnostic values of US. No meta-analysis could be done, but a weighted pooled sensitivity and specificity were calculated to 73% [61% - 83%] and 79% [68% - 88%], respectively.

**Comparison of the Imaging Modalities**

PET/CT had the highest diagnostic odds ratio of 20 [11 – 38], highest sensitivity and highest specificity compared to the other imaging modalities. Only four of the studies compared different imaging modalities on the same population, which were too few to do a direct comparison based on network meta-analysis. Meta-regression analysis on sub-groups showed no statistically significant superiority in joint diagnostic performance of either imaging modality. However, PET/CT showed significantly higher sensitivity compared to both CT (p-value = 0.01) and MRI (p-value < 0.001). See Figure 4 for the results of the meta-regression analyses using imaging type as covariate.

**Heterogeneity and Publication Bias**

The heterogeneity analyses showed an $I^2$ of 96, 81, and 88 for CT, MRI, and PET/CT, respectively. This was interpreted as substantial heterogeneity. The threshold effect was estimated to 0.27 for CT, 1.00 for MRI, and 1.00 for PET/MRI. For the studies on CT, meta-regression showed heterogeneity caused by publication year (p-value = 0.05), study quality (p-value = 0.02), and diagnostic criteria (p-value = 0.01). The heterogeneity among PET/CT studies could not be explained by the explored parameters in the meta-regression (no statistically significant coefficients). Due to few studies, no meta-regression could be done to explore heterogeneity among MRI and US accuracy values.

Deek’s funnel plot asymmetry test showed no sign of publication bias (p-value = 0.97).
Discussion

Summary of Main Findings
This is the first systematic review and meta-analysis on all imaging modalities for detection of ENE in HNSCC. By performing this study, we aimed to compare the diagnostic values among used imaging modalities to establish clinical guidance in the choice of optimal imaging for ENE detection. We found no significant difference in the overall ability of CT, MRI, and PET/CT to diagnose ENE. However, PET/CT had a significantly higher sensitivity than CT and MRI. Only two studies examined the diagnostic accuracy of US [10,27] making it impossible to include the modality in the meta-analysis. Currently, no studies on the diagnostic value of PET/MRI exist.

During the last decade, use of PET/CT in the diagnostic HNSCC work-up has increased dramatically. The modality combines high anatomical detailing and metabolic cell information. Recent studies have proved PET/CT superior to conventional imaging in the diagnostic assessment of HNSCC patients [37,38]. PET/CT has a sensitivity of 77% to 96% and a specificity of 82% to 100% to detect cervical lymph node metastases [39-41]. In comparison, the reported sensitivity and specificity for CT is ~77% and ~72%, respectively, whereas for MRI it is 72% and 81% [42]; and US 81% and 64% [43].

Our results on ENE detection were in line with these reports. We showed PET/CT to have the highest sensitivity (80% [76% - 84%]) compared with CT (76% [67% - 82%]) and MRI (72% [64% - 79%]). Contrary, a meta-analysis conducted by Su et al. from 2015 on the prediction of ENE in head and neck cancer showed no advantage of PET/CT compared to the other modalities [44]. However, only two PET/CT studies were included in the review [11,45]. Furthermore, Su et al. concluded that CT seemed to have a lower sensitivity (77% [70% - 82%]) compared with MRI (85% [80% - 89%]) [44]. In our study, we did not find any differences in the diagnostic values between CT and MRI. In fact, all diagnostic parameters except the specificity were higher for CT compared with MRI. A reason for the differing results may be the wider inclusion criteria used by Su et al. They included two additional Japanese studies in their
meta-analysis of MRI, which involved patients with nasopharyngeal carcinoma [46,47]. The two studies showed a very high MRI sensitivity of 100% [96% - 100%] and 96% [80% - 100], respectively, and accounted for 51% of the pooled sample size in the meta-analysis. Our meta-analysis included a recent study from 2020 that found the sensitivity of MRI to be of only 75% [66% - 83%], opposing the accuracy of MRI [24].

**Limitations**

No filters were used in our search strategy and only one paper was excluded due to lack of language capabilities (Chinese). The literature search was comprehensive and carried out with assistance from a specialized librarian, minimizing potential reporting bias. An obvious strength of our study was the large sample size with the inclusion of 3391 units in the meta-analysis.

A point of concern was the different observation units used in the included studies (i.e. lymph nodes, cervical regions, necks, or patients). However, we did not find any effect modification caused by the different observation units, when we explored sources of heterogeneity. Studies that based their analyses only on the largest suspicious lymph nodes may, however, be associated with selection bias. Large lymph nodes may harbor more aggressiveness, and consequently show more pronounced ENE, which are easier detected by the pathologist compared with smaller lymph nodes. Furthermore, it is known from previous studies that a major part of ENE is found in small lymph nodes [48]. By excluding these smaller lymph nodes in the analysis, diagnostic accuracy may be overestimated. On the other hand, it can be difficult to correlate small imaging-suspicious lymph nodes with the correct ones in the surgical specimens. Many of the included studies dealt with this issue by using neck level, neck side or patient as unit rather than the individual lymph node.

Our comparative meta-analysis on the different imaging modalities was limited by the lack of adequate studies with head-to-head comparisons. As recommended in the *Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy* [36], we addressed this issue by doing an indirect comparison of the
imaging modalities in a bivariate regression model with the imaging types as covariates. Although this method was limited by the heterogeneity in index test thresholds, we assessed it to be better than a simple univariate analysis.

Applicability of Findings

In terms of demographic features with an overweight of men above 60 years old, the studies were considered to cover the right population. Our review only included HNSCC patients with oropharyngeal, hypopharyngeal, laryngeal, and oral cavity cancers. Moreover, patients with squamous cell carcinoma metastases in lymph nodes with unknown primary tumor were also considered having HNSCC. These cancer sites were chosen with the clinical setting in mind, since surgical management with neck dissection in many of these cases is considered as the recommended treatment. All studies specified the T-site for all patients, except from three studies [14,23,25]. Although, these could diminish the external validity of the meta-analysis, we did not exclude them, since the studies did not mention inclusion of non-HNSCC patients. In terms of severity, almost all the studies specified the T- and N- stage. Both patients with high-stage and low-stage HNSCC were represented.

Variation in the use of imaging modalities was managed by pooling the meta-analysis individually for each imaging. However, there were some concerns regarding the variation in the diagnostic criteria used within each group. For instance, all PET/CT studies used different values of SUVmax (ranging from 2.3 to 10). This caused a threshold effect in the meta-analysis, which could be a possible explanation for the substantial heterogeneity. Threshold effect is a source of heterogeneity in diagnostic meta-analyses; the higher sensitivity (and lower thresholds), the lower specificity. We used the best diagnostic values reported in the studies. Thus, the threshold effect might be the reason for the significantly higher sensitivity of PET/CT, but non-higher diagnostic accuracy overall compared with the other modalities. However, when we introduced SUVmax as a continuous variable in the regression model, we did not find any statistically significant impact of the covariate.
Another topic of interest in the applicability of our findings is the histopathological definition of ENE. This is a controversial scientific discussion among pathologists, and to this day, no standardized definition exists. A high variability in the prevalence of ENE in HNSCC patients [1,49], and the poor inter- and intra-rater reliability among pathologists [50] indicate the need for a histopathological definition that pathologists agree on. In this review, most of the included studies had high risk of bias in the domain of the reference standard (table 2) due to lack of specifying the histopathological definition.

**Implications**

Based on this review, our recommendations do not alter the use of standard imaging in the work-up of patients with HNSCC. PET/CT showed good potential with a higher sensitivity. The combination of PET and CT seems advantageous and may replace the use of CT and MRI. However, further research with prospective head-to-head comparisons of PET/CT with other imaging modalities is warranted.

**Conclusion**

There was no significant difference in the ability of CT, MRI, and PET/CT to diagnose histopathological ENE, except that PET/CT had a significantly higher sensitivity than CT and MRI. There is a lack of head-to-head comparisons of relevant imaging modalities with regard to diagnosing ENE in HNSCC patients.

**Funding & Ethics**

_Funding:_ This work was supported by the University of Southern Denmark; the Region of Southern Denmark; Odense University Hospital; and the Harboe fund. The sponsors had no involvement in study design, data collection, analysis, interpretation of date, writing of the report, or decision to submit the article for publication.
Ethical approval: This article does not contain any studies with human participants performed by any of the authors

Acknowledgements

The authors thank Mette Brandt Eriksen, scientific librarian at the University of Southern Denmark, for helping with the development of the search strategy, and medical doctor Musa Yasin Kaya from Aarhus University Hospital for helping with the translation of Turkish articles in the review process.
References


Figure captions

Figure 1: Flow diagram of study selection

Figure 2: Forrest plot sorted by imaging modality. Studies are represented twice if data for two observers are available in the studies.

CT = computer tomography, PET = positron emission tomography with CT, MRI = magnetic resonance imaging, US = ultrasound

Figure 3: SROC for CT, MRI & PET/CT. Each study is represented as a circle. The 95% prediction contour is only illustrated for CT due to lack of enough MRI and PET/CT studies.

CT = computer tomography, PET = positron emission tomography, MRI = magnetic resonance imaging, SENS = sensitivity, SPEC = specificity, AUC = area under the curve

Figure 4: Meta-regression with subgroup analyses of different imaging modalities. Sensitivity and specificity to diagnose ENE are compared between modalities.

CT = computer tomography, PET = positron emission tomography, MRI = magnetic resonance imaging, vs. = versus
Table I: Characteristics of included studies. Assessors are given by the number of radiologists/observers in the study.

<table>
<thead>
<tr>
<th>Study</th>
<th>Year</th>
<th>Country</th>
<th>Study Type</th>
<th>Patients</th>
<th>Male (%)</th>
<th>Median Age (range)</th>
<th>T-Sites (n)</th>
<th>T-Stage (n)</th>
<th>N-Stage (n)</th>
<th>HPV Status (%)</th>
<th>Imaging</th>
<th>Threshold (Unit)</th>
<th>Sample Size (Unit)</th>
<th>Assessors (n)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aiken [6]</td>
<td>2015</td>
<td>USA</td>
<td>Retrospective</td>
<td>111</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>CT</td>
<td>Overall</td>
<td>111 (Patients)</td>
<td>1</td>
</tr>
<tr>
<td>Chun [7]</td>
<td>2016</td>
<td>Korea</td>
<td>Retrospective</td>
<td>89</td>
<td>80 (90)</td>
<td>62 (32 - 91) Larynx (89)</td>
<td>N/A</td>
<td>T1 (7)</td>
<td>T2 (25)</td>
<td>N0 (52) N1 (11) N2 (26)</td>
<td>N/A</td>
<td>PET/CT</td>
<td>SUV-max 2.8</td>
<td>62 (Regions)</td>
</tr>
<tr>
<td>Dequanter [8]</td>
<td>2015</td>
<td>Belgium</td>
<td>Retrospective</td>
<td>54</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>T2 (10)</td>
<td>T3 (6)</td>
<td>N/A</td>
<td>PET/CT</td>
<td>SUV-max 4.2</td>
<td>54 (Patients)</td>
<td>1</td>
</tr>
<tr>
<td>Geltzeiler [9]</td>
<td>2017</td>
<td>USA</td>
<td>Retrospective</td>
<td>100</td>
<td>80 (80)</td>
<td>N/A Oropharynx (100)</td>
<td>N/A</td>
<td>T0 (5)</td>
<td>T1 (45)</td>
<td>N/A</td>
<td>CT</td>
<td>Overall</td>
<td>70 (Patients)</td>
<td>1</td>
</tr>
<tr>
<td>Ishii [10]</td>
<td>2004</td>
<td>Japan</td>
<td>Retrospective</td>
<td>47</td>
<td>30 (63)</td>
<td>57 (27 - 83) CCO (47)</td>
<td>N/A</td>
<td>T1 (13)</td>
<td>T2 (20)</td>
<td>N/A</td>
<td>CT</td>
<td>US</td>
<td>Overall</td>
<td>48 (Necks)</td>
</tr>
<tr>
<td>Joo [12]</td>
<td>2013</td>
<td>Korea</td>
<td>Retrospective</td>
<td>80</td>
<td>55 (69)</td>
<td>54 (23 - 83) CCO (80)</td>
<td>N/A</td>
<td>T1 (29)</td>
<td>T2 (33)</td>
<td>N0 (45) N1 (11) N2 (24)</td>
<td>N/A</td>
<td>PET/CT</td>
<td>SUV-max 2.3</td>
<td>71 (Regions)</td>
</tr>
<tr>
<td>Joo [13]</td>
<td>2013</td>
<td>Korea</td>
<td>Retrospective</td>
<td>57</td>
<td>55 (96)</td>
<td>61 (35 - 80) Hypopharynx (57)</td>
<td>N/A</td>
<td>T1 (5)</td>
<td>T2 (25)</td>
<td>N0 (12) N1 (9) N2 (34) N3 (2)</td>
<td>N/A</td>
<td>PET/CT</td>
<td>SUV-max 2.7</td>
<td>86 (Regions)</td>
</tr>
<tr>
<td>Joo [11]</td>
<td>2013</td>
<td>Korea</td>
<td>Retrospective</td>
<td>78</td>
<td>74 (95)</td>
<td>59 (43 - 82) Oropharynx (78)</td>
<td>N/A</td>
<td>T1 (21)</td>
<td>T2 (41)</td>
<td>N0 (17) N1 (13) N2 (48) HPV+ (36)</td>
<td>PET/CT</td>
<td>SUV-max 3.9</td>
<td>106 (Regions)</td>
<td>1</td>
</tr>
<tr>
<td>Kanakamedala</td>
<td>2014</td>
<td>USA</td>
<td>Retrospective</td>
<td>77</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>CT</td>
<td>Necrosis</td>
<td>77 (Patients)</td>
<td>1</td>
</tr>
<tr>
<td>King [15]</td>
<td>2004</td>
<td>Hong Kong</td>
<td>Prospective</td>
<td>17</td>
<td>16 (94)</td>
<td>62 (50 - 85) Hypopharynx (7) Hypopharynx / Oropharynx (2) Larynx (2) CCO (6)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>CT</td>
<td>MRI</td>
<td>Overall</td>
<td>51 (Lymph nodes)</td>
</tr>
<tr>
<td>Lee [16]</td>
<td>2015</td>
<td>Korea</td>
<td>Prospective</td>
<td>186</td>
<td>138 (74)</td>
<td>64 (28 - 91) CCO (90) Oropharynx (43) Larynx (32) Hypopharynx (21)</td>
<td>N/A</td>
<td>T1 (50)</td>
<td>T2 (63)</td>
<td>N0 (71) N1 (36) N2 (76) N3 (3)</td>
<td>N/A</td>
<td>PET/CT</td>
<td>SUV-max 4.9</td>
<td>Overall 263 (Necks)</td>
</tr>
<tr>
<td>Study</td>
<td>Country</td>
<td>Type</td>
<td>Year</td>
<td>Patients</td>
<td>T1</td>
<td>T2</td>
<td>T3</td>
<td>T4</td>
<td>N1</td>
<td>N2</td>
<td>N3</td>
<td>N/A</td>
<td>Imaging</td>
<td>Other</td>
</tr>
<tr>
<td>-------</td>
<td>---------</td>
<td>------</td>
<td>------</td>
<td>----------</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>----</td>
<td>-----</td>
<td>--------</td>
<td>-------</td>
</tr>
<tr>
<td>Lodder [18]</td>
<td>Netherlands</td>
<td>Retrospective</td>
<td>2015</td>
<td>12 (83)</td>
<td>61 (50 - 75)</td>
<td>CCO (2)</td>
<td>Oropharynx (2)</td>
<td>Hypopharynx (1)</td>
<td>CUP (7)</td>
<td>N/A</td>
<td>N/A</td>
<td>PET/CT</td>
<td>SUV-max 12</td>
<td>18 (Lymph nodes)</td>
</tr>
<tr>
<td>Lodder [17]</td>
<td>Netherlands</td>
<td>Retrospective</td>
<td>2013</td>
<td>39 (25 (64))</td>
<td>63 (46 - 85)</td>
<td>CCO (23)</td>
<td>Oropharynx (5)</td>
<td>Hypopharynx (1)</td>
<td>CUP (10)</td>
<td>T1 (8)</td>
<td>T2 (10)</td>
<td>T3 (5)</td>
<td>T4 (6)</td>
<td>N1 (5)</td>
</tr>
<tr>
<td>Maxwell [19]</td>
<td>USA</td>
<td>Retrospective</td>
<td>2015</td>
<td>65 (60 (92))</td>
<td>56 (N/A)</td>
<td>Oropharynx (55)</td>
<td>Non-oropharynx (7)</td>
<td>CUP (3)</td>
<td>T1 (31)</td>
<td>T2 (18)</td>
<td>T3 (9)</td>
<td>T4 (4)</td>
<td>N1 (13)</td>
<td>N2 (50)</td>
</tr>
<tr>
<td>Noor [20]</td>
<td>Australia</td>
<td>Retrospective</td>
<td>2019</td>
<td>80 (68 (85))</td>
<td>57 (39 - 85)</td>
<td>Oropharynx (79)</td>
<td>CUP (1)</td>
<td>T1 (26)</td>
<td>T2 (41)</td>
<td>T3 (9)</td>
<td>T4 (3)</td>
<td>N1 (60)</td>
<td>N2 (20)</td>
<td>p16+ (100)</td>
</tr>
<tr>
<td>Patel [21]</td>
<td>USA</td>
<td>Retrospective</td>
<td>2018</td>
<td>27 (27 (100))</td>
<td>57 (N/A)</td>
<td>Oropharynx (27)</td>
<td>CUP (1)</td>
<td>T1/T2 (26)</td>
<td>T3 (1)</td>
<td>T4 (1)</td>
<td>N/A</td>
<td>N/A</td>
<td>p16+ (100)</td>
<td>CT</td>
</tr>
<tr>
<td>Randall [22]</td>
<td>Canada</td>
<td>Retrospective</td>
<td>2015</td>
<td>40 (29 (73))</td>
<td>N/A</td>
<td>CCO (40)</td>
<td>T1 (5)</td>
<td>T2 (19)</td>
<td>T3 (4)</td>
<td>T4 (12)</td>
<td>N1 (15)</td>
<td>N2 (25)</td>
<td>N3 (1)</td>
<td>N/A</td>
</tr>
<tr>
<td>Sharma [23]</td>
<td>India</td>
<td>Prospective</td>
<td>2017</td>
<td>30 (24 (80))</td>
<td>53 (N/A)</td>
<td>CCO (17)</td>
<td>Larynx (2)</td>
<td>Unknown (8)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>CT</td>
<td>Necrosis</td>
<td>30 (Patients)</td>
</tr>
<tr>
<td>Sheppard [24]</td>
<td>Switzerland</td>
<td>Retrospective</td>
<td>2020</td>
<td>186 (134 (72))</td>
<td>64 (31 - 86)</td>
<td>CCO (78)</td>
<td>Oropharynx (64)</td>
<td>Hypopharynx (15)</td>
<td>Larynx (12)</td>
<td>CUP (17)</td>
<td>T0(17)</td>
<td>T1(30)</td>
<td>T2(68)</td>
<td>T3(33)</td>
</tr>
<tr>
<td>Souter [25]</td>
<td>New Zealand</td>
<td>Retrospective</td>
<td>2009</td>
<td>124</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>CT</td>
<td>Overall</td>
<td>127 (Necks)</td>
</tr>
<tr>
<td>Steinkamp [28]</td>
<td>Germany</td>
<td>Prospective</td>
<td>1999</td>
<td>165 (135 (82))</td>
<td>58 (N/A)</td>
<td>Larynx (58)</td>
<td>Hypopharynx (30)</td>
<td>Oropharynx (21)</td>
<td>CCO (48)</td>
<td>CUP (8)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>CT</td>
</tr>
<tr>
<td>Steinkamp [26]</td>
<td>Germany</td>
<td>Prospective</td>
<td>2002</td>
<td>110 (91 (83))</td>
<td>58 (N/A)</td>
<td>Larynx (37)</td>
<td>Hypopharynx (22)</td>
<td>Oropharynx (16)</td>
<td>CCO (30)</td>
<td>CUP (5)</td>
<td>N/A</td>
<td>N/A</td>
<td>N/A</td>
<td>MRI</td>
</tr>
<tr>
<td>Study</td>
<td>Year</td>
<td>Location</td>
<td>Study Type</td>
<td>Cases</td>
<td>Cases (Total)</td>
<td>Cases (Primary Site)</td>
<td>Imaging</td>
<td>SUV - max</td>
<td>Cases (Surgical Site)</td>
<td>Imaging</td>
<td>SUV - max</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------</td>
<td>----------</td>
<td>------------</td>
<td>-------</td>
<td>--------------</td>
<td>---------------------</td>
<td>--------</td>
<td>----------</td>
<td>---------------------</td>
<td>--------</td>
<td>----------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Steinkamp [27]</td>
<td>2003</td>
<td>Germany</td>
<td>Prospective</td>
<td>110</td>
<td>91 (83)</td>
<td>Hypopharynx (22)</td>
<td>US</td>
<td>Indistinct boundaries</td>
<td>97 (Necks)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Oropharynx (16)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CCO (30)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Unknown (5)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toya [29]</td>
<td>2020</td>
<td>Japan</td>
<td>Retrospective</td>
<td>94</td>
<td>N/A</td>
<td>CCO (58)</td>
<td>N/A</td>
<td>PET/CT</td>
<td>SUV-max 3.0</td>
<td>566 (Regions)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Oropharynx (6)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hypopharynx (20)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Larynx (16)</td>
<td>N/A</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zoumalan [30]</td>
<td>2010</td>
<td>USA</td>
<td>Retrospective</td>
<td>17</td>
<td>17 (100)</td>
<td>Larynx (4)</td>
<td>T1 (1)</td>
<td>CT</td>
<td>Necrosis</td>
<td>61 (Lymph nodes)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CCO (6)</td>
<td>T2 (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Hypopharynx (3)</td>
<td>T3 (7)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Oropharynx (3)</td>
<td>T4 (1)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CUP (1)</td>
<td>N1 (4)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N2 (10)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>N3 (3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

CCO = cancer of the oral cavity, CUP = cancer of unknown primary tumor, CT = computer tomography, PET = positron emission tomography, MRI = magnetic resonance imaging, US = ultrasound,
Table 2: Risk of bias. Overall quality is based on the risk of bias assessment for each study in the domains of Patient, Index test, Reference Standard, and Flow and Timing.

<table>
<thead>
<tr>
<th>Study ID</th>
<th>Patient</th>
<th>Index Test</th>
<th>Reference Standard</th>
<th>Flow and Timing</th>
<th>Overall Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aiken 2015 [6]</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Chun 2016 [7]</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>High</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Dequanter 2015 [8]</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Ishii 2004 [10]</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Joo 2013 [12]</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Joo 2013 [13]</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Kanakamedala 2014 [14]</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
<td>Unclear</td>
<td>Acceptable</td>
</tr>
<tr>
<td>King 2004 [15]</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Lee 2015 [16]</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Lodder 2013 [17]</td>
<td>Low</td>
<td>High</td>
<td>Unclear</td>
<td>Low</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Lodder 2015 [18]</td>
<td>Unclear</td>
<td>High</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Maxwell 2015 [19]</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Noor 2019 [20]</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Patel 2018 [21]</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Randall 2015 [22]</td>
<td>Low</td>
<td>Unclear</td>
<td>High</td>
<td>Unclear</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Sharma 2017 [23]</td>
<td>High</td>
<td>Unclear</td>
<td>High</td>
<td>Low</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Sheppard 2020 [24]</td>
<td>Low</td>
<td>Low</td>
<td>Unclear</td>
<td>Low</td>
<td>High</td>
</tr>
<tr>
<td>Souter 2009 [25]</td>
<td>Low</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Steinkamp 2002 [26]</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Steinkamp 2003 [27]</td>
<td>Unclear</td>
<td>Low</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Steinkamp 1999 [28]</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Unclear</td>
<td>Acceptable</td>
</tr>
<tr>
<td>Toya 2020 [29]</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td>Acceptable</td>
</tr>
</tbody>
</table>
Figure 1: Flow diagram of study selection.

- Records identified through databases (n = 620)
  - Duplicates excluded (n = 144)
    - Records screened (n = 476)
      - Records excluded (n = 424)
        - Full-text articles assessed for eligibility (n = 52)
          - Full-text articles excluded (n = 27)
            - Wrong patient population (n = 13)
            - Wrong outcomes (n = 13)
            - Not included language (n = 1)
        - Qualitative synthesis: Full-text studies (n = 25)
          - Studies included in quantitative synthesis (meta-analysis) (n = 25)
Figure 2: Forrest plot sorted by imaging modality. Studies are represented twice if data for two observers are available in the studies.

CT = computer tomography, PET = positron emission tomography with CT, MRI = magnetic resonance imaging, US = ultrasound.
Figure 3: SROC for CT, MRI & PET/CT. Each study is represented as a circle. The 95% prediction contour is only illustrated for CT due to lack of enough MRI and PET/CT studies.

CT = computer tomography, PET = positron emission tomography, MRI = magnetic resonance imaging, SENS = sensitivity, SPEC = specificity, AUC = area under the curve
Figure 4: Meta-regression with subgroup analyses of different imaging modalities. Sensitivity and specificity to diagnose ENE are compared between modalities.

CT = computer tomography, PET = positron emission tomography, MRI = magnetic resonance imaging, vs. = versus