Radiology: Imaging Cancer

ORIGINAL RESEARCH

Interrater Reliability of NI-RADS on Posttreatment PET/Contrast-enhanced CT Scans in Head and Neck Squamous Cell Carcinoma

Derek Hsu, MD • Tanya J. Rath, MD • Barton F. Branstetter IV, MD, FACR • Yoshimi Anzai, MD, MPH • C. Douglas Phillips, MD • Amy F. Juliano, MD • Kristine M. Mosier, DMD, PhD • Michael P. Bazylewicz, MD • Stan M. Poliashenko, MD • Matthew H. Kulzer, MD • Patricia A. Rhyner, MD, FACR • Benjamin Risk, PhD • Richard H. Wiggins, MD • Ashley H. Aiken, MD

From the Department of Radiology, Emory University School of Medicine, 1364 Clifton Rd NE, Room BG03, Atlanta, GA 30322 (D.H., A.H.A.); Department of Neuroradiology, Mayo Clinic, Phoenix, Ariz (T.J.R.); Departments of Radiology and Otolaryngology, University of Pittsburgh School of Medicine, Pittsburgh, Pa (B.F.B.); Department of Radiology and Imaging Sciences, University of Utah Health Sciences Center, Salt Lake City, Utah (Y.A., R.H.W.); Department of Neuroradiology, Weill Cornell Imaging at New York–Presbyterian, New York, NY (C.D.P.); Department of Radiology, Massachusetts Eye and Ear, Harvard Medical School, Boston, Mass (A.F.J.); Department of Radiology and Imaging Sciences, Indiana University School of Medicine, Indianapolis, Ind (K.M.K.); Department of Radiology, University of Vermont Medical Center, Burlington, Vt (M.P.B.); Radiology Imaging Associates, Englewood, Colo (S.M.P.); Department of Radiology, Mayo Clinic, Jacksonville, Fla (M.H.K.); Department of Biostatistics and Bioinformatics, Emory University Rollins School of Public Health, Atlanta, Ga (P.A.R.); and Department of Radiology and Imaging Sciences, Emory University Hospital, Atlanta, Ga (B.R.). Received September 16, 2020; revision requested November 11; revision received March 16, 2021; accepted April 1. **Address correspondence to** D.H. (e-mail: *derek.evan.hsu@emory.edu*).

Authors declared no funding for this work. Conflicts of interest are listed at the end of this article.

Radiology: Imaging Cancer 2021; 3(3):e200131 • https://doi.org/10.1148/rycan.2021200131 • Content codes: CT HN OI

Purpose: To evaluate the interrater reliability among radiologists examining posttreatment head and neck squamous cell carcinoma (HNSCC) fluorodeoxyglucose PET/contrast-enhanced CT (CECT) scans using Neck Imaging Reporting and Data System (NI-RADS).

Materials and Methods: In this retrospective study, images in 80 patients with HNSCC who underwent posttreatment surveillance PET/CECT and immediate prior comparison CECT or PET/CECT (from June 2014 to July 2016) were uploaded to the American College of Radiology's cloud-based website, Cortex. Eight radiologists from seven institutions with variable NI-RADS experience independently evaluated each case and assigned an appropriate prose description and NI-RADS category for the primary site and the neck site. Five of these individuals were experienced readers (> 5 years of experience), and three were novices (< 5 years of experience). In total, 640 lexicon-based and NI-RADS categories were assigned to lesions among the 80 included patients by the eight radiologists. Light generalization of Cohen κ for interrater reliability was performed.

Results: Of the 80 included patients (mean age, 63 years ± 10 [standard deviation]), there were 58 men (73%); 60 patients had stage IV HNSCC (75%), and the most common tumor location was oropharynx (n = 32; 40%). Light κ for lexicon was 0.30 (95% CI: 0.23, 0.36) at the primary site and 0.31 (95% CI: 0.24, 0.37) at the neck site. Light κ for NI-RADS category was 0.55 (95% CI: 0.46, 0.63) at the primary site and 0.60 (95% CI: 0.48, 0.69) at the neck site. Percent agreement between lexicon and correlative NI-RADS category was 84.4% (540 of 640) at the primary site and 92.6% (593 of 640) at the neck site. There was no significant difference in interobserver agreement among the experienced versus novice raters.

Conclusion: Moderate agreement was achieved among eight radiologists using NI-RADS at posttreatment HNSCC surveillance imaging.

Supplemental material is available for this article.

In contrast to the traditional narrative radiology reports, a standardized radiology reporting system clarifies the degree of suspicion on the presence or absence of disease and directly links to subsequent management of patients. Furthermore, the Reporting and Data Systems allow for data collection and mining to better determine accuracy, interrater agreement, effect on management, and outcomes.

Created in 2016, the Neck Imaging Reporting and Data System (NI-RADS) is a standardized classification system with management recommendations linked to each category, which is based on the level of suspicion for the presence of tumor at head and neck squamous cell carcinoma (HNSCC) surveillance imaging after treatment. NI-RADS categories, ranging from 0 through 4, are assigned to the primary and neck (nodal) sites independently to convey the level of suspicion for disease at each site: 0, incomplete; 1, no evidence of recurrence; 2, low suspicion; 3, high suspicion; and 4, definitive recurrence. By replacing free-text narrative reports with standardized lexicon and a numerical category assignment, communication and further recommendations are clear for both the referring clinician and patient (1,2). Several recent reviews outline the practical application of NI-RADS and its improvement at head and neck cancer surveillance imaging (1–4).

PET/contrast-enhanced CT (CECT) has become an integral component in HNSCC surveillance (5–8). NI-RADS was initially developed for use with PET/CECT, and its performance has been examined in several published

Abbreviations

CECT = contrast-enhanced CT, DICOM = Digital Imaging and Communications in Medicine, HNSCC = head and neck squamous cell carcinoma, NI-RADS = Neck Imaging Reporting and Data System, RCMS = Radiology Case Management System

Summary

Eight radiologists from seven institutions achieved moderate interrater reliability using the standardized Neck Imaging Reporting and Data System in posttreatment head and neck squamous cell carcinoma.

Key Points

- Light κ for prose description was 0.30 at the primary site and 0.31 at the neck site, while Light κ for Neck Imaging Reporting and Data System (NI-RADS) category was 0.55 at the primary site and 0.60 at the neck site.
- Mismatch between chosen prose descriptions and NI-RADS category highlights an opportunity for education and the need for further training and standardization.
- Moderate interrater reliability using NI-RADS was achieved among radiologists with minimal NI-RADS experience, specifically with only three of the eight included radiologists using it in clinical practice.

Keywords

CT, PET/CT, Head/Neck, Neck, Neoplasms-Primary, Observer Performance

studies. Higher NI-RADS categories (ie, 3 and 4) at the baseline posttreatment PET/CECT and subsequent surveillance studies are all strongly associated with increased risk of treatment failure in patients with HNSCC. There is an increased rate of residual or recurrent disease with each increase in NI-RADS category, with rates of residual and recurrent disease reported between 3.8% and 4.3% for NI-RADS 1, 9%–17% for NI-RADS 2, 42%–59% for NI-RADS 3, and 100% for NI-RADS 4 (9–12).

While initial data have established the value of NI-RADS, interrater agreement among radiologists at different institutions is a topic to be investigated. Furthermore, the interrater agreement in the interpretation of the complicated posttreatment head and neck CECT has not been established, as its study has largely been precluded by the lack of standardized reporting. The development of NI-RADS provides the opportunity to study and improve interrater agreement and therefore improve our collective added value for head and neck cancer. Previously described interrater agreement between two radiologists from the same institution with prior experience using NI-RADS was very good, with a κ statistic of 0.82 (9). A subsequent study demonstrated moderate agreement among four radiologists from the same institution, with a κ statistic of 0.48 at the primary site and 0.50 at the neck site (13).

The purpose of this study was to examine the interrater reliability among radiologists from multiple different institutions examining posttreatment HNSCC PET/CECT scans using prose description and NI-RADS. To add value for patients with head and neck cancer in the posttreatment period, it is critical to understand the variation in interpretation and management on the basis of imaging findings. NI-RADS is a tool that now allows us to gather baseline data and generate educational tools to standardize radiologic interpretations.

Materials and Methods

This multi-institutional retrospective study was approved by the institutional review board at each respective institution and was in compliance with the Health Insurance Portability and Accountability Act. The requirement for written informed consent was waived.

Patient Selection and Case Creation

In this study, only patients with HNSCC who underwent pretreatment PET/CT or CECT and posttreatment PET/CECT between June 2014 and July 2016 from two large academic centers were included. All MR images were excluded. In total, 80 patients with HNSCC were selected from a database of posttreatment PET/CECT scans, with 88% (70 of 80) of patients from one institution, which were published on a prior study that examined patient clinical outcomes (11). This current study differs, as it analyzes interobserver agreement between radiologists interpreting imaging studies and includes additional patients from another institution, representing 12% (10 of 80) of the cohort.

Each of the 80 patients had a short history and two imaging studies, a pretreatment PET/CT or CECT to assess the extent of disease and a posttreatment PET/CT, which typically occurs between 8 to 12 weeks after treatment. A pretreatment examination was included to assess the primary tumor's original size and extent and the presence of any nodal disease. A short history provided raters with basic demographic information, including patient age, sex, tumor type and location, staging, and treatment history. A total of 306 individual examinations were uploaded, including a total of 148 fused PET/CT scans, 152 CECT scans, and six whole-body CT scans. Of these cases, 35 studies were from outside hospitals (35 of 306; 11.4%). Only axial images were uploaded for each case; no reformats were provided.

The patients' prior pretreatment and posttreatment examinations were uploaded online into cases in conjunction with the American College of Radiology through their cloud-based content manager, Radiology Case Management System (RCMS), and Cortex websites, which allow users to store, manage, and distribute radiologic content. All Digital Imaging and Communications in Medicine (DICOM) files were exported from the picture archiving and communication system and scrubbed of all identifying information by using OsiriX version 9.0.2 (Pixmeo). All cases were assigned a separate random numeric identifier and uploaded to RCMS. Cases were randomized in order by using a random number generator prior to upload on Cortex. DICOM images were displayed on Cortex through NilRead version 4.2.21.90144 (Hyland Software).

Image Acquisition

Fluorodeoxyglucose PET and CECT scans were performed with General Electric Discovery PET/CT scanners (GE Healthcare). Patients fasted for 4–6 hours prior to the scan. If serum glucose concentrations immediately prior to fluorodeoxyglucose administration were greater than 200 mg/dL, scanning was deferred. Combined PET/CT images from the skull vertex through the midthigh were obtained 50–60 minutes

after intravenous administration of 10–20 mCi of fluorodeoxyglucose, dosed by body weight. Helical non–contrast material–enhanced CT scans from the vertex through midthigh were performed before PET for attenuation correction and anatomic localization. Diagnostic CECT images were obtained either after a 45-second delay after administration of 125 mL of intravenous contrast agent or after a 90-second delay with split bolus technique of 110 mL of intravenous contrast agent. The contrast agent that was used was iopamidol (Isovue 370; Bracco). CT examinations followed one of two scan parameters: *(a)* 120–130 kV (peak); variable milliampere; pitch, 1.5–2; collimation, 3.75 mm or *(b)* 120 kV (peak); smart milliampere with a noise index of 13.78; pitch, 0.984; gantry rotation, 0.7 second; field of view, 25 cm. Scan parameters for pretreatment examinations from outside hospitals were not obtained.

Raters and Instructions

Eight radiologists with varying degrees of clinical experience (B.F.B., 19 years; Y.A., 19 years; C.D.P., 25 years; A.F.J., 14 years; K.M.M., 25 years; M.P.B., 3 years; S.M.P., < 1 year; M.H.K., < 1 year) from seven institutions nationwide were selected for the study. Five were experienced radiologists who were Certificate of Added Qualifications–certified neuroradiologists with more than 5 years of experience. Three novice radiologists included one general radiologist and two neuroradiology fellows. Radiologists had a variable degree of experience with NI-RADS. All participants were invited via e-mail and provided with a general overview of the study, a Cortex internet link to access cases, and a tutorial on navigating Cortex and viewing images on NilRead. Additionally, all readers were provided with an optional online teaching file consisting of a brief background about NI-RADS and practice cases (*https://www.acr.org/-/media/ACR/Files/RADS/NI-RADS/NIRADS-Atlas.pdf*).

Image Analysis and Questions

Cases were accessed through any PC, Mac, or Linux computer using any of the following compatible internet browsers: Google Chrome, Mozilla Firefox, Microsoft Internet Explorer, or Apple Safari. Each radiologist independently reviewed case images on a personal computer. Only axial images were provided to each rater. A subjective analysis of none versus mild versus intense fluorodeoxyglucose uptake was determined instead of a strict standard uptake value, as these data do not improve diagnostic accuracy for disease after treatment for HNSCC (5,14,15).

Subsequently, raters answered four multiple-choice questions (Appendix E1 [supplement]). The first two questions pertained to the primary site and the remaining two to the neck (nodal) site. Neck (nodal) assessment was based on the most suspicious-appearing node and not reflective of the total number of abnormal nodes. Questions 1 and 3 asked raters to select the most accurate prose description or lexicon for any visualized abnormality and were created to map to specific NI-RADS categories and the lexicon defined by the American College of Radiology categories descriptor table (Appendix E2 [supplement]). Questions 2 and 4 asked raters to assign the appropriate NI-RADS category at the primary site and neck site, respectively. Questions could be answered in any order and changed prior to final submission. No time restriction was enforced.

Power and Data Analysis

In determining the number of total patients required, we expected an average κ of approximately 0.8 and planned to demonstrate that it is higher than 0.6 by using a two-sided size test, with a significance level of .05 (16). With 75 patients split evenly across NI-RADS categories 1, 2, and combined 3 and 4, we would have at least 80% statistical power to demonstrate that a single κ coefficient is greater than 0.6. In practice, the frequencies of the NI-RADS categories were approximately 37%, 38%, and 25%, for NI-RADS 1, 2, and combined 3 and 4, respectively; updating with these numbers, n = 82 would achieve 80% power. By increasing the number of raters, this improves statistical power, making these estimates conservative.

Descriptive statistics were collated. Interrater agreement for prose descriptions (ie, lexicon responses) was assessed using unweighted Light generalization of Cohen κ measure of interrater reliability, which is based on the mean of the pairwise Cohen κ measures of agreement between the eight raters on the scoring of the 80 patients (17,18). NI-RADS category was treated as ordinal data with three levels (categories 1, combined 2a and 2b, and combined 3 and 4), and thus interrater agreement was assessed using Light κ with squared weights. The 95% CIs were calculated using 5000 bootstrap samples. NI-RADS 2a and 2b were combined, as they both represent low suspicion category, with the only distinction being the location of abnormality: mucosal versus submucosal. NI-RADS 3 and 4 were combined, as they both represent high-risk lesions for tumor recurrence. The κ statistic was interpreted using six different categories: less than 0, less than chance agreement; 0.01–0.20, slight agreement; 0.21–0.40, fair agreement; 0.41–0.60, moderate agreement; 0.61–0.80, substantial agreement; and 0.81–0.99, almost perfect agreement (19). All statistical analyses were performed on R version 3.6.1 (R Foundation for Statistical Computing) using the package irr (20).

Results

Patient and Questionnaire Overview

Of our 80 patients (mean age, 63 years ± 10), a majority of patients were men (73%; 58 of 80). The most common tumor stage was stage IV (75%; 60 of 80), and the most common location was the oropharynx (40%; 32 of 80). Half of the patients underwent chemoradiation therapy alone (50%; 40 of 80), and the other half underwent surgical treatment with or without adjuvant chemoradiation therapy (50%; 40 of 80). The average time interval between treatment completion and posttreatment imaging was 13.2 weeks ± 2.9. Table 1 shows patient and lesion characteristics. All four multiple-choice questions for each of the 80 patients were answered by the eight radiologists (2560 of 2560 questions, 100%).

Interobserver Responses on Questionnaire

Raters' prose description (ie, lexicon) responses from questions 1 and 3 were in the range of fair agreement at both the primary and neck sites. Light κ for lexicon was 0.30 (95% CI: 0.23, 0.36) at the primary site and 0.31 (95% CI: 0.24, 0.37) at the neck site (Table 2).

Raters' NI-RADS category assignments from questions 2 and 4 were in the range of moderate agreement at both the primary and neck sites. Weighted Light κ for NI-RADS category was 0.55 (95% CI: 0.46, 0.63) at the primary site and 0.60 (95% CI: 0.48, 0.69) at the neck site (Table 2).

Experience Level and Interobserver Agreement

There was no significant difference in interobserver agreement among experienced raters versus novice raters (Table 2). Light κ for lexicon at the primary site for experienced raters was 0.29 (95% CI: 0.22, 0.36) and for novice raters was 0.30 (95% CI: 0.19, 0.41). Light κ for lexicon at the neck site for experienced raters was 0.29 (95% CI: 0.22, 0.36), while for novice raters it was 0.32 (95% CI: 0.22, 0.41). Weighted Light κ for NI-RADS category at the primary site for experienced raters was 0.57 (95% CI: 0.47, 0.66) and for novice raters was 0.51 (95% CI: 0.36, 0.64). Weighted Light κ for NI-RADS at the neck site for experienced raters was 0.56 (95% CI: 0.44, 0.66), while for novice raters it was 0.67 (95% CI: 0.53, 0.79).

Agreement between Prose Description and NI-RADS Category

The prose description or lexicon used in multiple-choice questions 1 and 3 each has a specific NI-RADS category correlate. Raters' lexicon selections at the primary and neck sites were automatically mapped to the correct NI-RADS category through a mapping key and then compared with the raters' selected NI-RADS category (Fig 1). At the primary site, the percent agreement of the prose description and the correlative NI-RADS category was 84.4% (540 of 640), with a mismatch percentage of 15.6% (100 of 640). At the neck site, the percent agreement was 92.6% (593 of 640), with a mismatch percentage of 7.3% (47 of 640). Among our raters, the highest mismatch percentage was 37.5% (30 of 80) at the primary site and 27.5% (22 of 80) at the neck site (Table 3). Figures 2 and 3 were two exemplary cases that demonstrate discrepancies between our raters and highlight difficulties encountered with differentiating between NI-RADS 2a and 3.

Table 1: Patient and Lesion Characteristics

Characteristic	Value
No. of patients	80
No. of men	58 (73%)
Mean age ± standard deviation (y)	63 ± 10
Location	
Hypopharynx	1 (1%)
Larynx	23 (29%)
Nasopharynx	5 (6%)
Oral cavity	17 (21%)
Oropharynx	32 (40%)
Skin	1 (1%)
Unknown	1 (13%)
Treatment	
CRT	40 (50%)
Surgery with or without CRT	40 (50%)
Stage	
I	6 (8%)
II	3 (4%)
III	10 (13%)
IVA	55 (69%)
IVB	5 (6%)
Unknown	1 (1%)
NI-RADS category	
1	
Primary	35 (44%)
Neck	58 (73%)
2a	
Primary	26 (33%)
Neck	9 (11%)
2b	
Primary	2 (2%)
Neck	0 (0%)
3	
Primary	13 (16%)
Neck	11 (14%)
4	
Primary	4 (5%)
Neck	2 (2%)

Note.—CRT = chemoradiation therapy, NI-RADS = Neck Imaging Reporting and Data System.

Discussion

Interpretation of head and neck imaging, particularly posttreatment head and neck cancer, is very challenging (21). In fact, NI-RADS is the first multi-institutional standardized template to report imaging findings in patients with head and neck cancer. Recent studies have demonstrated the association between NI-RADS category and likelihood of treatment failure (9–12). NI-RADS has also facilitated evaluation of the prognostic value of specific radiologic findings at posttreatment HNSCC surveillance imaging. This study aimed to leverage the standardized categories used in NI-RADS to study interrater reliability in interpretation. Specifically, we wanted to evaluate agreement among radiologists from different institutions across the nation, who may be exposed to different practice patterns.

We found that interrater agreement among eight radiologists was fair when using prose descriptions (ie, lexicon) at both the primary and neck sites, with a Light κ of 0.30 and 0.31, respectively. It is not surprising that the interrater reliability even among subspecialists using prose description and impressions is only fair,

Table 2: Light Generalization of Cohen κ for Prose Description and NI-RADS Category

Site and Reader Experience Level	Prose Description	NI-RADS Category
All readers		
Primary	0.30 (0.23, 0.36)	0.55 (0.46, 0.63)
Neck	0.31 (0.24, 0.37)	0.60 (0.48, 0.69)
Experienced vs novice readers		
Primary		
Experienced	0.29 (0.22, 0.36)	0.57 (0.47, 0.66)
Novice	0.30 (0.19, 0.41)	0.51 (0.36, 0.64)
Neck		
Experienced	0.29 (0.22, 0.36)	0.56 (0.44, 0.66)
Novice	0.32 (0.22, 0.41)	0.67 (0.53, 0.79)

Note.—Values are shown with 95% CIs in parentheses. Experienced readers were those who had Certificate of Added Qualifications and more than 5 years of experience, while those with less than 5 years of experience were considered as novice. NI-RADS = Neck Imaging Reporting and Data System.

A

Primary		NI-RADS Description				
		1	2a	2b	3	4
NI-RADS Category	1	*233*	1	0	0	0
	2a	35	*178*	1	0	0
	2b	0	4	*21*	*5*	0
	3	0	33	2	*84*	8
	4	0	1	0	10	*24*

B

Neck		NI-RADS Description			
		1	2	3	4
NI-RADS Category	1	*365*	3	0	0
	2	14	*111*	0	0
	3	3	9	*89*	0
	4	0	1	17	*28*

Figure 1: Prose description versus Neck Imaging Reporting and Data System (NI-RADS) category, *A*, at the primary site and, *B*, at the neck site.

Table 3: Discordance Rates among Individual Raters between Prose Descriptions and NI-RADS Category

Rater	Experience	Primary (%)	Neck (%)
1	E	23 (18/80)	14 (11/80)
2	N	10 (8/80)	1 (1/80)
3	E	18 (14/80)	5 (4/80)
4	E	38 (30/80)	10 (8/80)
5	N	6 (5/80)	0 (0/80)
6	E	28 (22/80)	28 (22/80)
7	E	4 (3/80)	0 (0/80)
8	N	0 (0/80)	1 (1/80)
Overall	NA	16 (100/640)	7 (47/640)

Note.—E = experienced (had Certificate of Added Qualifications and more than 5 years of experience), N = novice (less than 5 years of experience), NA = not applicable, NI-RADS = Neck Imaging Reporting and Data System.

especially in the interpretation of these complex postsurgical and postradiation studies (22,23). A 2017 study showed a similar lack of interobserver agreement in the interpretation of lumbar spine MRI (24). This is an opportunity for head and neck radiologists to lead the way to improve standardization. Approximately 65% of patients with head and neck cancer will have treatment failure after definitive treatment (23). A recent study shows that nearly a third of these treatment failures will be asymptomatic patients with recurrences identified by imaging alone (25). It is therefore critical that we clearly communicate our findings and standardize our level of suspicion on the basis of specific findings that have either been shown to be more likely after treatment (NI-RADS 2) or more likely treatment failure (NI-RADS 3).

When using NI-RADS categories, interrater agreement increased to moderate. Weighted Light κ was 0.55 at the primary site and 0.60 at the neck site. While the κ statistic may be partly increased with the use of weighted κ for ordinal data in the NI-RADS category, in contrast to categorical data for prose descriptions, NI-RADS category agreement was higher than prose description at both the primary and neck sites. We hypothesize that the discrete suspicion levels of NI-RADS encourage the radiologist to commit to an opinion that drives the next step in management, which has limited possibilities, namely: 1, no change with routine clinical observation; 2, closer follow-up or directed clinical examination; 3, biopsy and/or surgery; 4, palliation and/or salvage treatment. These management options map to NI-RADS 1, 2, 3, and 4, respectively. In other words, radiologists simply decide if the examination is definitely normal with expected posttreatment change (NI-RADS 1), suspicious for tumor recurrence or definitely abnormal (NI-RADS 3 or 4), or indeterminate as to whether an abnormality represents posttreatment change or tumor (NI-RADS 2). If radiologists are not highly suspicious that an abnormality is tumor, then the best course is often close imaging follow-up or directed visual examination and not biopsy. It is possible that the close connection with NI-RADS category with subsequent management paradoxically helps to clarify suspicion level for the radiologist and simplifies interpretation of the complex posttreatment neck. This in turn improves interobserver agreement, which is valuable as categories have predictive value (9–11). Nevertheless,

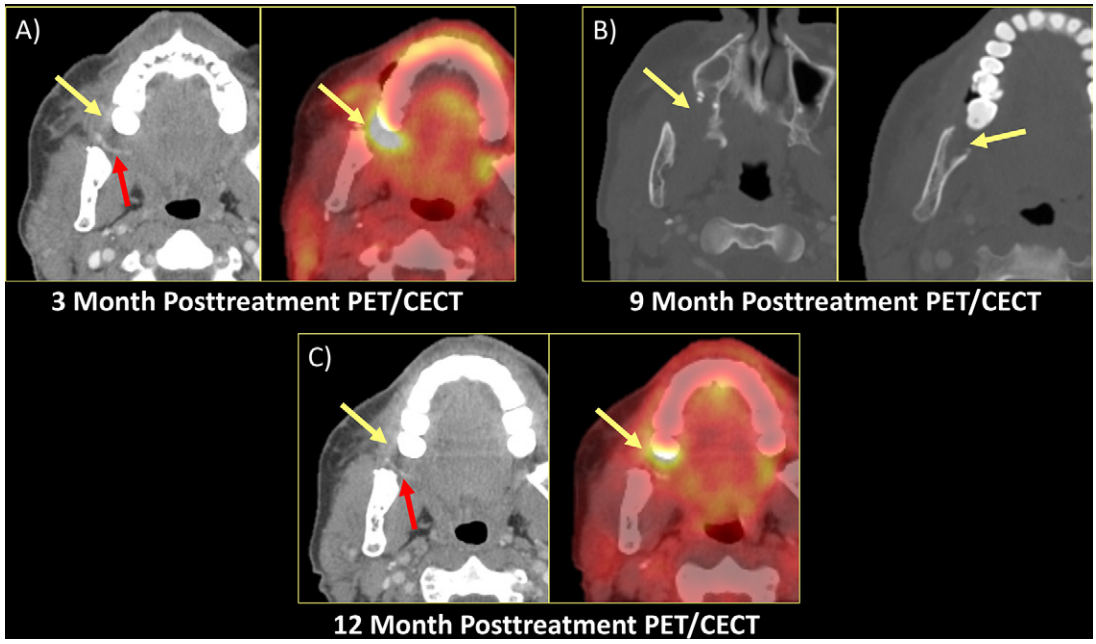

Figure 2: T2N0M0 squamous cell carcinoma of the right buccal mucosa treated with wide local excision, selective neck dissection, and right forearm free flap. *A,* At 3-month posttreatment contrast material–enhanced CT, there was a linear area of enhancement along the lateral aspect of the right first mandibular molar (yellow arrow). At PET, this area was associated with intense fluorodeoxyglucose uptake (yellow arrow). Of note, the curvilinear enhancing lesion represents the pedicle of the forearm free flap (red arrow). Clinical examination findings were not suggestive of tumor recurrence but felt to represent tumefactive radiation injury. *B,* A contrast-enhanced neck CT scan with bone window obtained 9 months after treatment demonstrates destructive osseous changes centered around the right maxillary tuberosity and right mandible (yellow arrow). CT-guided biopsy revealed fibrotic tissue with inflammation, consistent with osteoradionecrosis. *C,* At imaging 12 months after treatment, there is interval improvement in fluorodeoxyglucose uptake (yellow arrow). Half of the raters assigned the first posttreatment study a NI-RADS category 3 at the primary site. However, given its mucosal location for direct clinical examination, NI-RADS 2a is more appropriate. Red arrow indicates curvilinear enhancing lesion representing pedicle of forearm free flap.

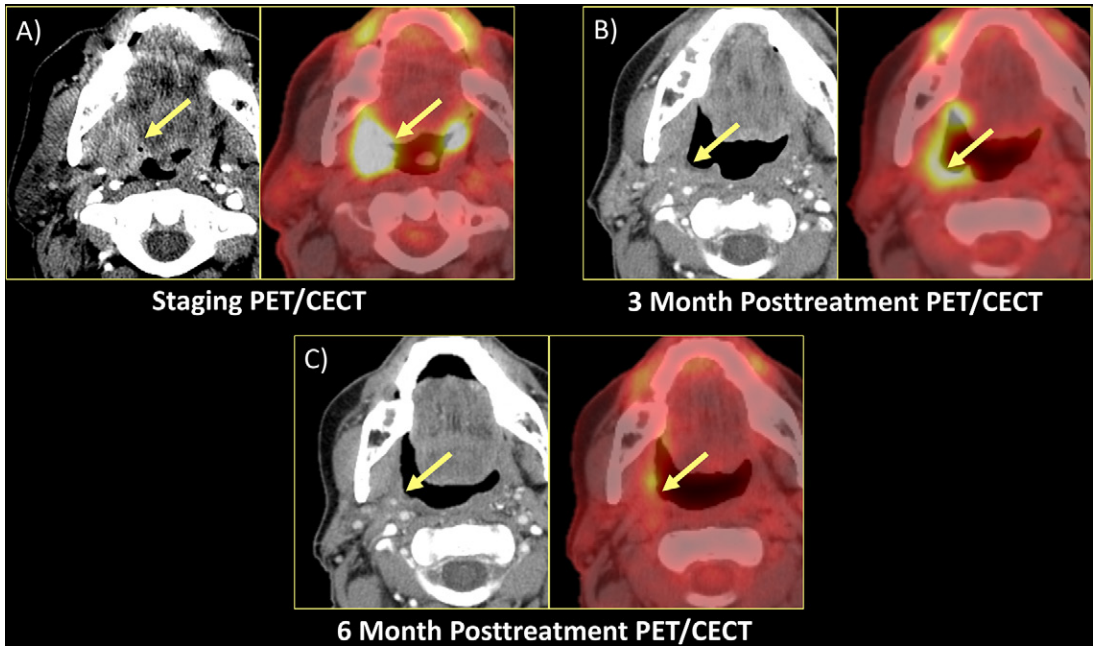

Figure 3: *A,* T4aN1M0 squamous cell carcinoma of the right palatine tonsil staging PET/contrast-enhanced CT (CECT). Patient was treated with chemoradiation therapy alone. *B,* At 3 months after treatment, PET/CECT scan demonstrated an enhancing soft-tissue mass along the superior margin of ulceration in the region of palatine tonsil with corresponding intense fluorodeoxyglucose uptake, concerning for residual tumor. However, upon direct visualization, the otolaryngologists suspected the imaging findings were radiation injury to soft tissues. A CT-guided biopsy revealed inflammatory cells and necrotic debris. *C,* On 6-month follow-up images, there is marked improvement in fluorodeoxyglucose uptake. This case was particularly difficult, with half of our raters assigning a NI-RADS category 2a and the others a 3, and became an index case to help define the 2a category. Immediate posttreatment PET/CECT scans with mucosal-based abnormalities are now typically assigned a NI-RADS category 2a, as many of these are tumefactive radiation injury.

our study highlights that there is still room for improvement, as the majority of the readers (five of eight) only had a brief tutorial on NI-RADS and at the time of the study had not incorporated NI-RADS into their clinical practice.

As previously described in two studies, interrater agreement from single institutions reached a κ statistic of 0.82 in 40 patients between two experienced readers and achieved a κ statistic of 0.48 at the primary site and 0.50 at the neck site among four readers (9,13). Our current study has several key differences. First, we included radiologists from multiple geographic regions and different institutions with variable experience with NI-RADS. Whereas previously, radiologists from the same institution were included. Second, our study included a larger number of patients when compared with Krieger et al (9). Third, the NI-RADS category distribution of the 40 previously selected patients used for the κ statistic is unclear in Krieger et al. However, when comparing the overall population, we have a higher percentage of abnormal scans when compared with these two studies. We chose to have a higher percentage of abnormal scans when compared with the prevalence in the disease population to prevent a prohibitively large cohort.

Accurate NI-RADS category assignment requires knowledge of the template lexicon, as each prose description used in our questions has an appropriate NI-RADS category correlate. There were a total of 147 instances of mismatch between the lexicon and NI-RADS category among our raters, 100 at the primary site and 47 at the neck site. Certain posttreatment changes led to most of the discrepancies. For example, diffuse, linear mucosal enhancement correlates with NI-RADS 1, but some readers assigned it NI-RADS category 2; or focal-mucosal enhancement, which correlates with NI-RADS 2a, was assigned NI-RADS category 3 by some raters (Fig 2). These mismatches highlight an opportunity for education such as appropriate lexicon and NI-RADS category assignment or specific exemplary cases to help delineate differences between challenging descriptions and categories, such as NI-RADS 2a versus NI-RADS 3 (Fig 3).

There were several limitations to our study. First, interpretation of the κ statistic varies among many researchers (26). We evaluated our κ statistic using one of the most common guidelines, as it provides easy interpretation (19). Additionally, while the κ statistic accounts for the number of categories, an increased number of categories may lower overall agreement (27). Second, raters were limited by constraints of an online browser interface instead of a diagnostic-grade picture archiving and communication system station and were provided only axial images without reformats. Third, our raters viewed a higher number of abnormal and higher-stage tumor cases when compared with published HNSCC study populations (9,11). In the setting of complex surgical and radiation treatments, immediate posttreatment surveillance imaging is challenging. With a higher incidence of patients with NI-RADS 3 and higher-stage tumors, we anticipated more variability with the inclusion of more advanced tumors. If lower-stage tumors were included, better interrater agreement may have been achieved. Last, a majority of our radiologists do not use NI-RADS in their daily practice and were using the template for the very first time in this study. The previously reported κ statistic of 0.82 likely is the result of the two neuroradiologists' extensive experience using NI-RADS, indicating the presence of a learning curve with NI-RADS (9). With further training, education, and experience with NI-RADS, we expect that interrater reliability will improve.

In conclusion, our study demonstrates moderate agreement among eight radiologists from multiple institutions using NI-RADS with posttreatment HNSCC PET/CECT scans. Light κ was 0.55 at the primary site and 0.60 at the neck site. Decreased agreement between prose descriptions and mismatch between lexicon and NI-RADS category highlight opportunities for further education and stress the need for further training and standardization.

Acknowledgments: We thank Dr Ajeet Mehta, Ms Lauren Attridge, and Mr Santosh Reddy for their diligent work and support on this project.

Author contributions: Guarantors of integrity of entire study, D.H., A.F.J., A.H.A.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, D.H., T.J.R., C.D.P., A.F.J., P.A.R., A.H.A.; clinical studies, D.H., T.J.R., B.F.B., Y.A., K.M.M., M.P.B., M.H.K., R.H.W.; statistical analysis, B.F.B., A.F.J., B.R., A.H.A.; and manuscript editing, D.H., T.J.R., B.F.B., Y.A., C.D.P., A.F.J., K.M.M., M.P.B., S.M.P., P.A.R., R.H.W., A.H.A.

Disclosures of Conflicts of Interest: **D.H.** disclosed no relevant relationships. **T.J.R.** disclosed no relevant relationships. **B.F.B.** disclosed no relevant relationships. **Y.A.** disclosed no relevant relationships. **C.D.P.** disclosed no relevant relationships. **A.F.J.** disclosed no relevant relationships. **K.M.M.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is an employee of Indiana University Health Physicians; author gave expert testimony in 2021 for Fumuso, Kelly, Farrell, Polin & Christesen; author's institution has grants/grants pending with National Institutes of Health (SPARC 3OTOD023847); author receives royalties from Elsevier for Statdx Amirsys. Other relationships: disclosed no relevant relationships. **M.P.B.** disclosed no relevant relationships. **S.M.P.** disclosed no relevant relationships. **M.H.K.** disclosed no relevant relationships. **P.A.R.** disclosed no relevant relationships. **B.R.** disclosed no relevant relationships. **R.H.W.** Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: author is employed by University of Utah Health Sciences Center. Other relationships: disclosed no relevant relationships. **A.H.A.** disclosed no relevant relationships.

References

1. Aiken AH, Rath TJ, Anzai Y, et al. ACR Neck Imaging Reporting and Data Systems (NI-RADS): A White Paper of the ACR NI-RADS Committee. J Am Coll Radiol 2018;15(8):1097–1108.
2. Juliano AF, Aiken AH. NI-RADS for head and neck cancer surveillance imaging: What, why, and how. Cancer Cytopathol 2020;128(3):166–170.
3. Aiken AH, Hudgins PA. Neck Imaging Reporting and Data System. Magn Reson Imaging Clin N Am 2018;26(1):51–62.
4. Hsu D, Juliano AF. Neck Imaging Reporting and Data System: Principles and Implementation. Neuroimaging Clin N Am 2020;30(3):369–377.
5. Ong SC, Schöder H, Lee NY, et al. Clinical utility of 18F-FDG PET/CT in assessing the neck after concurrent chemoradiotherapy for Locoregional advanced head and neck cancer. J Nucl Med 2008;49(4):532–540.
6. Koshkareva Y, Branstetter BF 4th, Gaughan JP, Ferris RL. Predictive accuracy of first post-treatment PET/CT in HPV-related oropharyngeal squamous cell carcinoma. Laryngoscope 2014;124(8):1843–1847.
7. McDermott M, Hughes M, Rath T, et al. Negative predictive value of surveillance PET/CT in head and neck squamous cell cancer. AJNR Am J Neuroradiol 2013;34(8):1632–1636.
8. Taghipour M, Sheikhbahaei S, Wray R, et al. FDG PET/CT in Patients With Head and Neck Squamous Cell Carcinoma After Primary Surgical Resection With or Without Chemoradiation Therapy. AJR Am J Roentgenol 2016;206(5):1093–1100.
9. Krieger DA, Hudgins PA, Nayak GK, et al. Initial Performance of NI-RADS to Predict Residual or Recurrent Head and Neck Squamous Cell Carcinoma. AJNR Am J Neuroradiol 2017;38(6):1193–1199.

10. Wangaryattawanich P, Branstetter BF 4th, Hughes M, Clump DA 2nd, Heron DE, Rath TJ. Negative Predictive Value of NI-RADS Category 2 in the First Posttreatment FDG-PET/CT in Head and Neck Squamous Cell Carcinoma. AJNR Am J Neuroradiol 2018;39(10):1884–1888.
11. Hsu D, Chokshi FH, Hudgins PA, et al. Predictive Value of First Posttreatment Imaging Using Standardized Reporting in Head and Neck Cancer. Otolaryngol Head Neck Surg 2019;161(6):978–985.
12. Wangaryattawanich P, Branstetter BF, Ly JD, Duvvuri U, Heron DE, Rath TJ. Positive Predictive Value of Neck Imaging Reporting and Data System Categories 3 and 4 Posttreatment FDG-PET/CT in Head and Neck Squamous Cell Carcinoma. AJNR Am J Neuroradiol 2020;41(6):1070–1075.
13. Elsholtz FHJ, Ro SR, Shnayien S, et al. Inter- and Intrareader Agreement of NI-RADS in the Interpretation of Surveillance Contrast-Enhanced CT after Treatment of Oral Cavity and Oropharyngeal Squamous Cell Carcinoma. AJNR Am J Neuroradiol 2020;41(5):859–865.
14. Leung AS, Rath TJ, Hughes MA, Kim S, Branstetter BF 4th. Optimal timing of first posttreatment FDG PET/CT in head and neck squamous cell carcinoma. Head Neck 2016;38(Suppl 1):E853–E858.
15. Ho AS, Tsao GJ, Chen FW, et al. Impact of positron emission tomography/computed tomography surveillance at 12 and 24 months for detecting head and neck cancer recurrence. Cancer 2013;119(7):1349–1356.
16. Flack VF, Afifi AA, Lachenbruch PA, Schouten HJA. Sample size determinations for the two rater kappa statistic. Psychometrika 1988;53(3):321–325.
17. Conger AJ. Integration and generalization of kappas for multiple raters. Psychol Bull 1980;88(2):322–328.
18. Light RJ. Measures of response agreement for qualitative data: Some generalizations and alternatives. Psychol Bull 1971;76(5):365–377.
19. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360–363.
20. irr: Various Coefficients of Interrater Reliability and Agreement. https://CRAN.R-project.org/package=irr. Published 2019. Accessed 2019.
21. Saito N, Nadgir RN, Nakahira M, et al. Posttreatment CT and MR imaging in head and neck cancer: what the radiologist needs to know. RadioGraphics 2012;32(5):1261–1282; discussion 1282–1284.
22. Argiris A, Karamouzis MV, Raben D, Ferris RL. Head and neck cancer. Lancet 2008;371(9625):1695–1709.
23. Chow LQM. Head and Neck Cancer. N Engl J Med 2020;382(1):60–72.
24. Herzog R, Elgort DR, Flanders AE, Moley PJ. Variability in diagnostic error rates of 10 MRI centers performing lumbar spine MRI examinations on the same patient within a 3-week period. Spine J 2017;17(4):554–561.
25. Gore A, Baugnon K, Beitler J, et al. Posttreatment Imaging in Patients with Head and Neck Cancer without Clinical Evidence of Recurrence: Should Surveillance Imaging Extend Beyond 6 Months? AJNR Am J Neuroradiol 2020;41(7):1238–1244.
26. Gwet KL. Agreement Coefficients for Nominal Ratings: A Review. In: Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters. 4th ed. Gaithersburg, Md: Advanced Analytics, 2014; 27–69.
27. Maclure M, Willett WC. Misinterpretation and misuse of the kappa statistic. Am J Epidemiol 1987;126(2):161–169.