**18F-FDG PET/CT Staging of Head and Neck Cancer: Interobserver Agreement and Accuracy—Results from Multicenter ACRIN 6685 Clinical Trial**

Rathan M. Subramaniam1,2, Fenghai M. Duan3, Justin Romanoff3, Jian Qin Yu4, Twyla Bartel5, Farrokh Dehdashti6, Charles M. Intenzo7, Lilja Solnes8, JoRean Sicks3, Brendan C. Stack Jr9, and Val J. Lowe10

1Otago Medical School, University of Otago, Dunedin, New Zealand; 2Duke University, Durham, North Carolina; 3School of Public Health, Brown University, Providence, Rhode Island; 4Fox Chase Cancer Center, Philadelphia, Pennsylvania; 5Global Advanced Imaging, Little Rock, Arkansas; 6Washington University School of Medicine, St Louis, Missouri; 7Thomas Jefferson University, Philadelphia, Pennsylvania; 8Johns Hopkins School of Medicine, Baltimore, Maryland; 9Southern Illinois School of Medicine, Springfield, Illinois; and 10Mayo Clinic, Rochester, Minnesota

To our knowledge, no prior multicenter clinical trial has reported interobserver agreement of 18F-FDG PET/CT scans for staging of clinical N0 neck in head and neck cancer. **Methods:** A total of 287 participants were recruited. For visual analysis, positive nodal uptake of 18F-FDG was defined as uptake visually greater than activity seen in the blood pool. **Results:** The negative predictive value of the 18F-FDG PET/CT for N0 clinical neck was 86% or above for visual assessment (95% CI, 86%–88%) for the 2 central readers and above 90% (95% CI, 90%–95%) for SUVmax, for central reads and site reads dichotomized at the optimal cutoff value of 1.8 and the prespecified cutoff value of 3.5, respectively. The $k$ coefficients between the 2 expert readers and between central reads and site reads varied between 0.53 and 0.78. **Conclusion:** The NPV of the 18F-FDG PET/CT for N0 clinical neck was 86% or above for visual assessment and above 90% for SUVmax cut points of 1.8 and 3.5 with moderate to substantial agreements. **Key Words:** oncology: head and neck; FDG PET/CT; head and neck cancer; staging

**J Nucl Med 2022; 63:1887–1890**
DOI: 10.2967/jnumed.122.263902

PET/CT with 18F-FDG is commonly used in clinical practice for management of head and neck squamous cell carcinoma patients including for staging, treatment assessment, and detecting recurrence and metastases (1–5). We previously reported on the primary results of ACRIN 6685 trial (ClinicalTrials.gov identifier: NCT00983697) (5,6). No prior multicenter study reported interobserver agreement for staging clinical N0 neck in head and neck cancer. In this post hoc analysis study, we report on the interobserver agreement among the readers interpreting the 18F-FDG PET/CT studies and their accuracy.

**MATERIALS AND METHODS**

**Patient Population**

As previously described, a total of 287 participants were recruited (Fig. 1) (5). A clinically N0 neck was defined as being free of palpable lymph nodes and with neck CT or MRI neck lymph node sizes of less than 1 and 1.5 cm for jugular digastric nodes (Ila), spinal accessory nodes (Iib), or submental-submandibular nodes (Ia and Ib) or showing a lack of central lymph node necrosis in nodes of any size (5).

**Imaging Procedure and Interpretation**

Imaging procedures and interpretation methods were previously described (5). PET/CT images were read at each study site by the reporting physician (i.e., site reads) and images were presented to a core reading panel of board-certified nuclear medicine or nuclear radiology certified physicians. There were 2 central readers: reader 1 and reader 2 (expert head and neck readers) who interpreted most of the PET/CT scans for the study. In addition, reader 3 and reader 4 (general readers) were used because central readers 1 and 2 were excluded from reading scans from their respective institutions and when adjudication was needed. A SUVmax was required for the hottest lymph node for each nodal basin recorded as indeterminate, probably malignant, or definitely malignant. The SUVmax calculation was performed using commercial software (version 5.2; MIM Software). For visual analysis, positive nodal uptake of 18F-FDG was defined as uptake visually greater than background and more than that activity seen in the blood pool (Fig. 2).

**Statistical Analysis**

The neck-level visual assessment 18F-FDG PET/CT scan result for each central reader, for the sites and for the central adjudicated read, was compared with the neck-level pathology result. The sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were calculated. Similar analyses were performed to compare the nodal basin SUVmax result (dichotomized at the optimal cutoff value of 1.8 (5) and the prespecified cutoff value of 3.5) with the nodal-level pathology. Cohen’s $k$ statistic was used to assess the agreement between the 2 expert readers (central readers 1 and 2) and the central readers and site reads. Because of data sparsity, agreement assessment for the 2 general readers (central readers 3 and 4) was not reported.

For all analyses, 95% CIs were calculated using the 2.5 and 97.5 percentiles of the multilevel bootstrap based on 10,000 resampled datasets (5). Analyses were performed using SAS software (version 9.4; SAS Institute) and R (version 4.0.4; R Foundation for Statistical Computing).

**RESULTS**

**Patient Demographics**

Patient characteristics are included in Supplemental Table 1 (supplemental materials are available at http://jnm.snmjournals.org).
which include data on enrolled patients and those who were included in this post hoc analyses.

**Visual Assessment**

There were 4 central readers: reader 1 and reader 2 (expert head and neck readers), and reader 3 and reader 4 (general readers). Readers 1, 2, 3, and 4 interpreted a total of 286, 273, 34, and 26 sides of necks, respectively. The site readers interpreted a total of 296 sides of neck. The sensitivity, specificity, PPV, and NPV of the visual assessment for the 2 expert central readers, the site readers, and the central adjudicated read are summarized in Table 1. The \( \kappa \) coefficients comparing reader 1 and reader 2, reader 1 and the central adjudicated read, reader 2 and the central adjudicated read, and the site reads and the central adjudicated read were 0.549 (95% CI: 0.431, 0.660), 0.756 (95% CI: 0.664, 0.837), 0.781 (95% CI: 0.696, 0.856), and 0.531 (95% CI: 0.421, 0.633), respectively.

**SUV\(_{\text{max}}\) Reads**

Readers 1, 2, 3, and 4 analyzed a total of 2,272, 2,171, 270, and 208 neck nodes measuring SUV\(_{\text{max}}\), respectively. The site readers analyzed a total of 2,385 neck nodes. The sensitivity, specificity, PPV, and NPV of SUV\(_{\text{max}}\) for the 2 expert readers and central adjudicated read are summarized in Table 2 for cut points 1.8 and 3.5. The \( \kappa \) statistics for measuring the agreement between the site SUV\(_{\text{max}}\) and the combined central SUV\(_{\text{max}}\) were 0.447 (95% CI: 0.363, 0.527) and 0.525 (95% CI: 0.382, 0.649), respectively, for SUV\(_{\text{max}}\) cut points of 1.8 and SUV\(_{\text{max}}\) 3.5. The \( \kappa \) coefficients for measuring the agreement between reader 1 and the combined central SUV\(_{\text{max}}\) were 0.818 (95% CI: 0.758, 0.870) and 0.751 (95% CI: 0.642, 0.839), respectively, for SUV\(_{\text{max}}\) cut points of 1.8 and SUV\(_{\text{max}}\) 3.5. The \( \kappa \) coefficients for measuring the agreement between reader 2 and the combined central SUV\(_{\text{max}}\) were 0.712 (95% CI: 0.640, 0.777) and 0.839 (95% CI: 0.741, 0.915), respectively, for SUV\(_{\text{max}}\) cut points 1.8 and 3.5.

**DISCUSSION**

The NPV of the \(^{18}\)F-FDG PET/CT for N0 clinical neck was 86% or above for visual assessment (95% CI, 86%–88%) for 2 expert central readers, and above 90% (95% CI, 90%–95%) for SUV\(_{\text{max}}\) cut points of 1.8 and 3.5 for the 2 expert readers and site reads. There was moderate to substantial agreement between readers. Increasing evidence supports the higher NPV of PET/CT to exclude nodal metastasis (5.7–9). In this study, we have provided evidence that multiple readers can achieve high NPV by visual assessment as well as by SUV\(_{\text{max}}\) analysis. This result has significant implications, especially managing the contralateral neck, as single-center studies have now reported on the outcome of patients managed with observation of PET-directed (negative) contralateral neck (10,11). The interreader reliability varied between moderate and substantial agreement in this study. Using the ACRIN 6685 standardized interpretation algorithm (visual assessment) may improve the reliability of interpretation more than subjective individual reader interpretation. It is important to note that there was moderate agreement between site readers and central readers, without any training for the site readers, which simulates day-to-day clinical practice. To our knowledge, there is no other baseline interpretation schema for neck nodal assessment using \(^{18}\)F-FDG PET/CT scans that has undergone interreader reliability assessment at a multicenter level. The standardized qualitative criteria (12), such as Hopkins criteria (2), NI-RADS (13), Deauville (14), and Porceddu (15), are for post-therapy settings. The interreader reliability for SUV\(_{\text{max}}\) readings between central and site readers appears lower than previously reported in single-center studies for interreader and intrarreader agreements (16,17), which is likely due to statistical reporting as a dichotomous (based on SUV\(_{\text{max}}\) cut points of 1.8 and 3.5) measure than a continuous measure.
One of the limitations of the ACRIN 6685 reads was that no detailed neck nodal level visual interpretation was performed though SUV<sub>max</sub> analysis was done. As the visual interpretation was recorded as side of the neck positive or negative for nodal metastasis, a global assessment was obtained. Another limitation for the SUV<sub>max</sub> inter-reader agreement is readers may have recorded SUV<sub>max</sub> of different lymph nodes at the same neck nodal level, which each reader considered positive and led to lower interreader agreement for SUV<sub>max</sub> than observed in single-center studies.

**CONCLUSION**

The NPV of the 18F-FDG PET/CT for N0 clinical neck was 86% or above for visual assessment (95% CI, 86%–88%) and above 90% (95% CI, 90%–95%) for SUV<sub>max</sub> cut points of 1.8 and 3.5. There is moderate to substantial agreement between central readers, between site reads and central adjudicated read, and central readers and central adjudicated read.

**DISCLOSURE**

ACRIN 6685 was supported by the National Cancer Institute through grants U01 CA079778, U01 CA080098, CA180820, and CA180794. No other potential conflict of interest relevant to this article was reported.

**KEY POINTS**

**QUESTION:** What is the NPV and reader reliability of 18F-FDG PET/CT for staging head and neck cancer with clinical N0 neck in a multicenter trial?

**PERTINENT FINDINGS:** The NPV of the 18F-FDG PET/CT for N0 clinical neck was 86% or above for visual assessment (95% CI, 86%–88%) and above 90% (95% CI, 90%–95%) for SUV<sub>max</sub> cut points of 1.8 and 3.5 for the 2 expert readers and site reads, with moderate to substantial agreement between all readers.

**IMPLICATIONS FOR PATIENT CARE:** 18F-FDG PET/CT has very high NPV for staging clinical N0 neck and has moderate to substantial interreader reliability, especially between site and central readers, which is important for day-to-day clinical practice.

**REFERENCES**


**TABLE 1**

<table>
<thead>
<tr>
<th>Reader</th>
<th>Sensitivity</th>
<th>Specificity</th>
<th>PPV</th>
<th>NPV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Expert H&amp;N reader 1</td>
<td>0.791</td>
<td>0.584</td>
<td>0.417</td>
<td>0.881</td>
</tr>
<tr>
<td>Expert H&amp;N reader 2</td>
<td>0.683</td>
<td>0.724</td>
<td>0.466</td>
<td>0.866</td>
</tr>
<tr>
<td>Central adjudicated read</td>
<td>0.740</td>
<td>0.644</td>
<td>0.443</td>
<td>0.866</td>
</tr>
<tr>
<td>Site read</td>
<td>0.700</td>
<td>0.699</td>
<td>0.471</td>
<td>0.859</td>
</tr>
</tbody>
</table>

H&N = head and neck.
Data in parentheses are 95% CIs.


### TABLE 2

<table>
<thead>
<tr>
<th>Diagnostic Test Statistics for the Dichotomized SUV_{\text{max}} Result Versus Pathology</th>
<th>1.8 cutoff value for SUV_{\text{max}}</th>
<th>3.5 cutoff value for SUV_{\text{max}}</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reader</strong></td>
<td><strong>Sensitivity</strong></td>
<td><strong>Specificity</strong></td>
</tr>
<tr>
<td>Expert H&amp;N scan reader 1</td>
<td>0.471</td>
<td>0.944</td>
</tr>
<tr>
<td>Expert H&amp;N scan reader 2</td>
<td>0.250</td>
<td>0.900</td>
</tr>
<tr>
<td>Combined central SUV_{\text{max}}</td>
<td>0.507</td>
<td>0.965</td>
</tr>
<tr>
<td>Site read</td>
<td>0.395</td>
<td>0.945</td>
</tr>
</tbody>
</table>

Data in parentheses are 95% CIs.