Oral Oncology 77 (2018) 105–110

Contents lists available at ScienceDirect

Oral Oncology

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

Reader performance in the ultrasonographic evaluation of oropharyngeal carcinoma

Farhoud Faraji[a,1], Ericka S. Padilla[b], Dana Blitz[b], Meghan B. Wenderoth[b], Ray G. Blanco[c], Satomi Kawamoto[b], Sheila Sheth[b], Ulrike M. Hamper[b], Carole Fakhry[a,*]

[a] *Department of Otolaryngology-Head and Neck Surgery, Johns Hopkins School of Medicine, Baltimore, MD, United States*
[b] *The Russell H. Morgan Department of Radiology and Radiological Sciences, Johns Hopkins Hospital, Baltimore, MD, United States*
[c] *Department of Surgery, Greater Baltimore Medical Center, Baltimore, MD, United States*

ARTICLE INFO

Keywords:
Head and neck
Squamous cell carcinoma
HNSCC
Oropharyngeal
OPSCC
Tonsil
Base of tongue
Ultrasound

ABSTRACT

Objective: To examine reader performance in evaluating oropharyngeal anatomy on ultrasonography.
Materials and methods: Ultrasound images of the oropharynx comprising normal and malignant anatomic variants were organized into slideshows. Slideshows were administered to 6 readers blinded to participant tumor status and with varying experience reading oropharyngeal sonograms. A training slideshow oriented readers to images of the oropharynx with and without malignant lesions. Readers then evaluated images in a test slideshow for tumor presence and marked orthogonal long and short dimensions of the tumor. Results were analyzed for accuracy, sensitivity, specificity, inter-reader agreement, and measurement error relative to prospectively-identified reference measurements.
Results: Eighty-seven percent of base of tongue (BOT) sonograms were identified correctly by a majority of readers. In identifying BOT tumors, median accuracy, sensitivity, specificity, and Fleiss's kappa were 79%, 73%, 85%, and 0.51, respectively. Median measurement error in the long and short axes for BOT tumors was −2.6% (range: −40% to 29%) and −2.6% (range: −56% to 156%), respectively. Eighty-four percent of palatine tonsil sonograms were identified correctly by a majority of readers. In identifying tonsil tumors, median accuracy, sensitivity, specificity, and Fleiss's kappa were 77%, 74%, 78%, and 0.41, respectively. Median measurement error in the long and short axes for tonsil tumors was 3.8% (range: −45% to 32%) and −6.5% (range: −83% to 42%), respectively.
Conclusions: Overall, US has clinically useful sensitivity for identification of oropharyngeal carcinoma among readers of diverse clinical backgrounds and experience. US may be useful for the evaluation of features such as tumor dimensions.

Introduction

Oropharyngeal carcinoma (OPC) is the most common head and neck cancer in North America [1] and is steadily increasing in incidence [2,3]. HPV-related OPC (HPV-OPC) arises primarily from the lymphoid-associated epithelia of the palatine tonsils and base of tongue (BOT) and exhibits a high rate of metastasis to cervical lymph nodes [4,5]. Patients with HPV-OPC often present with a persistently enlarged neck node, unaware of a small primary tumor present in the oropharynx [6]. HPV-OPCs are difficult to assess clinically due to their small size, and the anatomic topology of the oropharynx [6–11]. The anatomic challenges posed by the structure of the oropharynx often result in the need for examination under anesthesia with direct laryngoscopy for accurate assessment of tumor extent. Currently, no easily accessible, non-invasive method exists for the reliable clinical evaluation of oropharyngeal carcinoma.

Our group has investigated the potential role of transcervical ultrasonogaphy (US) in the evaluation of oropharyngeal tumors. US can be used to identify the primary tumor site of head and neck tumors undetectable by comprehensive clinical and imaging-based evaluation [12–14]. An optimized transcervical US protocol implementing anatomic landmarks for high-confidence visualization of oropharyngeal

* Corresponding author at: Johns Hopkins University School of Medicine, Otolaryngology-Head and Neck Surgery, 601 N. Caroline Street, Sixth Floor, Baltimore, MD 21287, United States.
E-mail addresses: ffaraji@slu.edu (F. Faraji), epadill4@jhmi.edu (E.S. Padilla), dblitz2@jhmi.edu (D. Blitz), mwendero13@gmail.com (M.B. Wenderoth), RBlanco@gbmc.org (R.G. Blanco), skawamo1@jhmi.edu (S. Kawamoto), ssheth@jhmi.edu (S. Sheth), umhamper@jhu.edu (U.M. Hamper), cfakhry@jhmi.edu (C. Fakhry).
[1] Present address: Saint Louis University School of Medicine, Saint Louis, MO, United States.

https://doi.org/10.1016/j.oraloncology.2017.12.012
Received 13 August 2017; Received in revised form 11 December 2017; Accepted 20 December 2017
1368-8375/

tumors was recently reported [15]. Given the learning curve of US we aimed to determine whether US readers blinded to participant tumor status can distinguish normal oropharynx from cancer-involved oropharynx. Determining whether independent US readers can differentiate oropharynx with and without malignancy is critical prior to further development of this technique. In this report, the performance of blinded US readers in identifying HPV-OPC from tonsil and BOT sonograms acquired by our optimized transcervical US protocol was examined.

Materials and methods

Study population

Twenty-four participants with biopsy-confirmed HPV-positive oropharyngeal carcinoma were prospectively enrolled for pretreatment transcervical ultrasonography of the oropharynx. Three control participants without oropharyngeal carcinoma were also enrolled. Ultrasonography was performed per standardized protocol for imaging oropharyngeal structures as previously described [15]. Procedures described in this study were approved by the Johns Hopkins Institutional Review Board.

Data collection: reader training and testing

Ultrasonography was performed by trained sonographers (M.B.W., E.S.P., and D.B.). The Philips iU22 or Philips EPIQ7 (Koninklijke Philips N.V., Amsterdam, Netherlands) ultrasound systems were used with the C5-1, C8-5, and X6-1 transducers. The C5-1 and C8-5 transducers were used to image the base of tongue. The C8-5 transducer was used to image the palatine tonsils. The X6-1 transducer was used for xPlane imaging, capturing two full-resolution planes (coronal and sagittal) simultaneously and displaying them in juxtaposed panels.

Prior to ultrasound imaging, sonographers reviewed each participant's clinical findings, including cross-sectional imaging, physical exam, fiberoptic laryngoscopy, and biopsy results. Gold standard reference tumor measurements were defined as the prospectively-measured tumor dimensions at the time of ultrasonography. Decisions for measurements were made by experienced sonographers in partnership with the lead study coordinator (F.F.) after confirmation of tumor dimensions and location on clinical imaging and other clinical evaluation.

Static ultrasound images that demonstrated a suspicious lesion consistent with clinical findings and tissue diagnosis were considered tumor images. Control images were captured from participants without oropharyngeal carcinoma (n = 3), from the contralateral tonsil of participants with tonsillar carcinoma (n = 7), from the tonsils of participant with BOT carcinoma (n = 5), from the BOT of participants with tonsillar carcinoma (n = 10), or from contralateral BOT of participants with BOT carcinoma (n = 12).

Given the predilection of HPV-OPC for the palatine and lingual tonsils, this study was focused upon normal palatine tonsils and lingual tonsils. Therefore, static ultrasound images of anatomically normal and tumor-involved BOT and palatine tonsil were selected for training and test slideshows. Each slide of training set initially presented an unlabeled ultrasound image. As readers advanced through the slideshow, labels appeared on the image to orient the reader to relevant anatomic structures. As labels appeared, they indicated image orientation (sagittal versus coronal), anatomic landmarks including boundaries of the normal BOT or tonsil, outlines of tumor extent (if present), and familiarized readers with testing procedures, including the identification and measurement of tumors. In the test set, image orientation and anatomic landmark labels were omitted and readers indicated whether a tumor was identified in each image. If a tumor was identified, readers measured the longest dimension of the tumor and its corresponding orthogonal short dimension. Unlabeled versions of images presented in the training set were included in the test set to compare reader accuracy for previously-encountered (training) and new images. All presented images were static and no video clips were included in the training or test sets, precluding dynamic interpretation by scanning through anatomic structures.

The BOT training set included 11 tumor and 5 control images. The BOT test set was comprised of 33 tumor and 58 control images. The tonsil training set included 4 tumor and 4 control images. The tonsil test set contained 31 tumor and 31 control images. In total, readers viewed 107 BOT and 78 tonsil images, 24 of which they previously encountered in annotated form in the training set.

Base of tongue and tonsil slideshows were administered to each of the 6 readers: 2 sonographers, 1 head and neck surgeon, and 3 radiologists. The 2 sonographers were trained in the standardized oropharyngeal ultrasonography protocol. Though neither of the sonographers had previously seen ultrasound images presented in the training or test sets, one sonographer had experience imaging and interpreting 11 oropharyngeal tumors while the other had imaged 2 oropharyngeal tumors. The head and neck surgeon had extensive experience evaluating oropharyngeal tumors with ultrasonography in clinic. The radiologists were experts in ultrasound; evaluating sonograms comprised significant portions of their respective practices. Regarding specific experience interpreting oropharyngeal US, one radiologist had moderate experience, while the other two had little to no prior experience reading oropharyngeal ultrasound images. The sonographers who evaluated images for this study did not participate in capturing any of the images used in this study. None of the readers were involved in the creation of the training or test sets.

Statistical analysis

Inter-reader reliability was determined by calculating Fleiss's kappa [16]. A proposed scale for interpreting the strength of inter-reader agreement suggests the following ranges for kappa values: < 0 (poor), 0–0.20 (slight), 0.21–0.4 (fair), 0.41–0.6 (moderate), 0.61–0.80 (substantial), 0.81–1.00 (near perfect) [17].

Images correctly indicated by readers as tumor that failed to accurately mark the lesion on the slide were scored as false negatives on accuracy, sensitivity, and inter-reader reliability analyses. Error in the measurement of long and short axes was defined as the percent difference in axis length as measured by the reader and the prospectively-measured dimensions during ultrasonography. For each image, average measurement error across all readers was calculated. Dimensions measured by readers that did not coincide with the pre-determined lesion on reference measurements were excluded from measurement error analysis. Boxplots were generated across all images to display the median error, interquartile range, and total range.

Bland-Altman analysis was performed to determine the relationship between tumor size by reference and reader measurements [18]. Reader measurement error was defined as the sum of the reference measurement and the absolute value of the mean error. In addition, percent measurement error was stratified by tumor category. Kruskal-Wallis H-test was employed to test for potential differences in the distribution of measurement error with respect to tumor category [19].

Results

Participant and tumor characteristics

The majority of the study population (n = 27) was men (n = 22, 82%) and white (n = 22, 82%) participants ranging from 26 to 81 years in age (Table 1). Biopsy confirmed subsite was BOT only (n = 12, 44%), palatine tonsil only (n = 8, 30%), and overlapping lesions of BOT and tonsil (n = 4, 15%). Eighty-three percent of tumors were cT1 (38%) or cT2 (46%). Three controls were included.