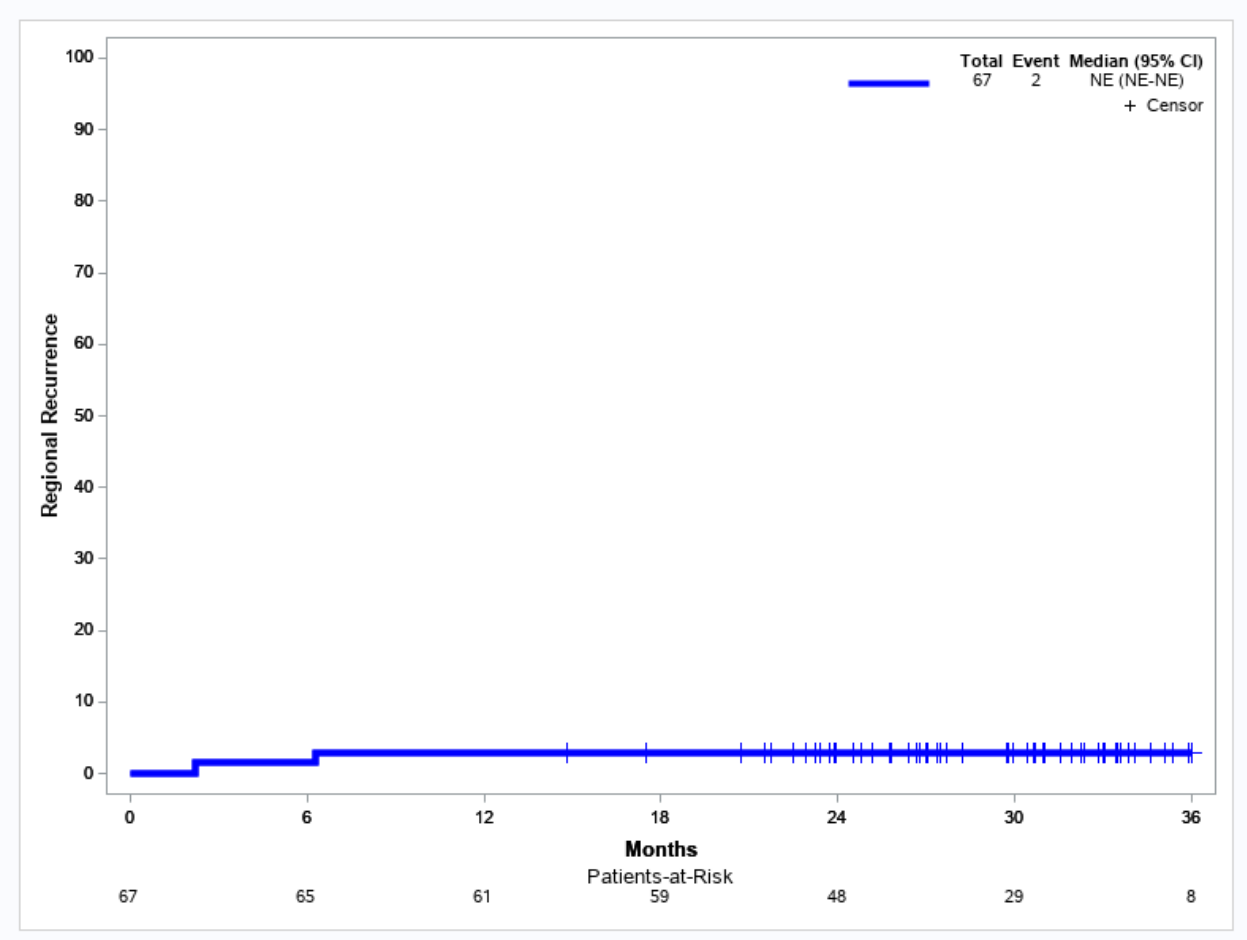
Supplementary Materials

Supplementary Figure 1. Cumulative incidence of regional recurrence in the entire cohort.



Supplementary Table 1. Representativeness of Study Participants

Cancer type	Head and neck squamous cell carcinoma
Sex	The ratio for human papillomavirus (HPV)-positive head and neck squamous cell carcinoma (HNSCC) incidence in males versus females is in the range of 3–6. (1)
Age	In a well-established study of oropharyngeal cancer, the median age at diagnosis for HPV-associated oropharyngeal cancer is approximately 57 years, while the median age of diagnosis for non-HPV associated oropharyngeal cancer was 61 years. (2)
Race/ethnicity	From 1992 through 2014, the overall incidence rate of HNSCC per 100,000 persons per year in the US was 11.2, ranging from 6.8 (Asian non-Hispanic) to 7.2 (Hispanic), to 12.2 (White non-Hispanic) to 14.3 (Black non-Hispanic). The comparative numbers for oropharyngeal squamous cell carcinoma were 3.4 (overall), 1.2 (Asian non-Hispanic), 1.9 (Hispanic), 4.0 (White non-Hispanic), and 4.4 (Black non-Hispanic). However, by 2014, the incidence rates for both any HNSCC and oropharyngeal cancer were highest in White individuals. (3)
Geography	In 2018, there were approximately 52,000 cases of head and neck cancer in the United States and approximately 10,000 deaths from the disease. Worldwide, there are 890,000 new cases and 450,000 deaths from head and neck cancer. (4)
Overall representativeness of this study	The median age of patients in our study was 62 years, with the population 79% male. In total, 72% of the cohort was comprised of oropharynx cancer. These numbers track with the general presentation of head and neck cancer in the country. However, the vast majority of patients (91%) were white, which to some extent limits generalizability of the results. However, it is unlikely that the probability of occult nodal disease differs by race.

1. Viens, L. J. et al. Human papillomavirus-associated cancers – United States, 2008–2012. *Morb. Mortal. Wkly Rep.* **65**, 661–666 (2016).
2. O'Sullivan, B et al. Development and validation of a staging system for HPV-related oropharyngeal cancer by the International Collaboration on Oropharyngeal cancer Network for Staging (ICON-S): a multicentre cohort study. *Lancet Oncol* 2016; 17:440-51.
3. Fakhry, C et al. Head and neck squamous cell cancers in the United States are rare and the risk now is higher among white individuals compared with black individuals. *Cancer* 2018; 124: 2125-33.
4. Chow, L.Q.M Head and neck cancer. N Engl J Med 2020; 382(1): 60-72.

Supplementary Methods

Dataset for lymph node malignancy prediction

In order to define a "ground truth" of the malignancy status of a lymph node, we used pathology results from patients with head and neck squamous cell carcinoma of the oropharynx, larynx and hypopharynx who underwent neck dissection from 8/2011 to 5/2018 at our institution. Preoperative CT and PET-CT from these patients were imported and fused in our contouring system (Velocity, Varian Medical Systems, Palo Alto, CA), and each individual lymph node was contoured and associated with its pathology result. Association of the preoperative imaging and pathologic diagnosis was facilitated by our surgical standard of labeling each lymph node station during resection; if a diagnosis was uncertain, that lymph node was not included for model generation.

We developed three separate AI models to predict a lymph node's malignancy status based on these contours. One model was created to analyze the simulation CT and fused PET imaging and included all patients with preoperative PET-CT. However, not every patient in the training set received a PET-CT prior to surgery, and the information within smaller nodes is presumably contained mostly within the CT data. Therefore, we also generated two disease-specific CT-only models (oropharynx and larynx/hypopharynx).

AI-based Lymph Node Malignancy Prediction Model

The lymph node malignancy prediction models integrate outputs from a multi-objective radiomics (MO-Radiomics) model that utilizes handcrafted imaging features and a convolutional neural network (CNN) that relies on learned features from PET and CT.

MO-radiomics model: The MO-radiomics model includes the following three key steps: 1) Quantitative imaging feature extraction from PET and CT images; 2) Predictive model construction and training using an iterative multi-objective immune algorithm; 3) Selection of an optimal solution. In MO-Radiomics, imaging features including intensity, texture, and geometric features were extracted from physician-contoured lymph nodes in PET and CT images. Intensity features include minimum, maximum, mean, stand deviation, sum, median, skewness, kurtosis, and variance. Geometry features include volume, major diameter, minor diameter, eccentricity, elongation orientation, bounding box volume, and perimeter. Texture features are based on 3D gray-level co-occurrence (GLCM) and gray level run-length (GLRL), which are extracted as follows: energy, entropy, correlation, contrast, texture variance, sum-mean, inertia, cluster shade, cluster prominence, homogeneity, max-probability, and inverse variance. A total of **257** features were extracted for PET and CT images. We used the support vector machine (SVM) to construct the predictive model with parameters denoted by $\alpha = [\alpha_1, \cdots, \alpha_M]$; M is the number of model parameters. Features extracted from PET, and CT are denoted by $\beta = [\beta_1, \cdots, \beta_N]$, where N is the number of features. The goal of the MO-radiomics model is to maximize sensitivity (f_{sen}) and specificity (f_{spe}) simultaneously to obtain the Pareto-optimal set: $f = \max_{\alpha,\beta}(f_{sen}, f_{spe})$. To solve the MO-radiomics optimization problem, we developed a multi-objective optimization algorithm[1]. During the model optimization, feature selection and model parameters training were performed simultaneously. The first phase of the optimization is to generate a Pareto-optimal solution set. The individual in all feasible solutions was sorted in descending order using a fast non-dominated sorting approach[2] according to the AUC of each solution and the final solution was selected with the highest AUC in MO-radiomics.

CNN-based model: For each lymph node, we use a cropped volume with dimensions of 64x64x48 in voxel size (equivalent to 32x32x24 mm^3) containing the lymph node from contrast-enhanced CT and PET-based imaging. This size was specifically selected to encompass the largest lymph node and its surrounding tissue. In this context, the term "surrounding tissues" refers to any tissues located within a 10-voxel expansion in all three dimensions surrounding the largest lymph node. Since all lymph nodes are contoured for MO-Radiomics modeling, the centroid of the contoured LN is used to define the center of the bounding box, which has a size of 64x64x48 voxels. Data augmentation was performed by a 3D rotation of [330°:30°] along a random axe of the three axes. The CNN architecture includes 12 convolutional layers, 2 max pooling layers, and 2 fully connected layers. Instead of using 2D convolutional kernels for image classification, our CNN uses 3D kernels in all convolutional layers. For CT-only model, the input of the network is a volumetric image. When we use PET and contrast-enhanced CT together, the input of the network contains two volumetric images, and each volumetric image serves as one channel of the input. The categorical cross entropy was the loss function.

Integrated model combining MO-Radiomics and CNN

Manually extracted features and automatically learned features can be complementary as demonstrated in several studies[3,4], including ours[5,6]. As such, a strategy that combined both handcrafted and learning models was used to predict the final lymph node malignancy in this study. The analytical evidential reasoning approach[1] was used to compute the final output probabilities for testing samples, by fusing the output probabilities generated by both the MO-Radiomics and CNN models. This approach involved combining belief degrees and weighting factors of the two models to generate basic probability assignments, which were then normalized

using a normalization constant to a range of [0,1]. Weighting factors of each model for fusion were calculated based on the model performance measured by AUC on the validation set.

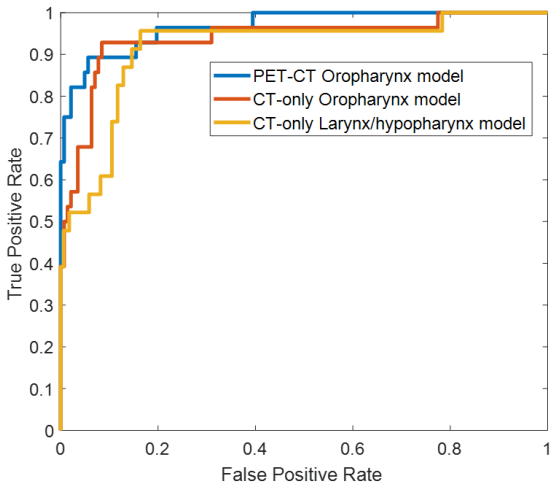
Performance of AI-based lymph node malignancy classification models

For the generation of the combined PET-CT and CT lymph node prediction model, a total of 791 lymph nodes from 129 patients were labeled as malignant or benign. Approximately 20% of the lymph nodes were malignant. The final PET-CT model was trained on lymph nodes from 80% of the patients; its performance on the remaining data (170 lymph nodes), measured via sensitivity, specificity, accuracy, and area under the receiver operating characteristic curve (AUROC), was 0.93, 0.85, 0.86, and 0.97, respectively. For the 146 (86% of total cohort) lymph nodes with cross-section diameter less than 17mm (primary criterion for AI-evaluated nodes in the trial), the model achieved 0.88, 0.86, 0.86, and 0.94 for sensitivity, specificity, accuracy and AUROC, respectively.

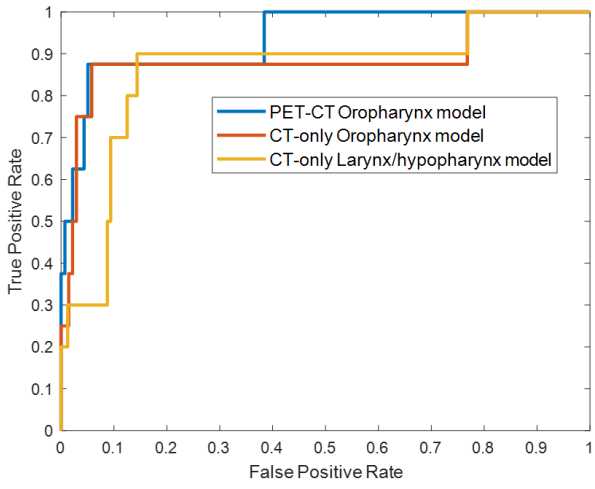
The CT-only oropharynx model used data from 78 patients and was trained and tuned in a similar manner as above using the same data, excluding the PET-CT imaging. The final sensitivity, specificity, accuracy, and AUROC were 0.93, 0.86, 0.87, and 0.94, respectively. For 146 (86%) lymph nodes with cross-sectional diameter less than 17 mm, the model achieved 0.88, 0.88, 0.88, and 0.88 for sensitivity, specificity, accuracy and AUROC, respectively. The CT-only larynx/hypopharynx model was initialized using the CT-only oropharynx model's weights as a baseline but was later fine-tuned with 51 patients with larynx/hypopharynx cancer who had preoperative CT imaging. A total of 386 lymph nodes were labeled, approximately 10% of which were malignant. Fifty percent of the larynx patients was used for training and hyperparameter selection, and 50% of patients (with 194 lymph nodes) were used for the final validation. The

final sensitivity, specificity, accuracy, and AUROC were 0.96, 0.84, 0.85, and 0.92, respectively. For the 170 (88%) lymph nodes with cross-section diameter less than 17mm, the model achieved 0.90, 0.86, 0.86, and 0.86 for sensitivity, specificity, accuracy and AUROC, respectively.

The combined AUROC performance of the entire cohort is shown in Figure 1(a) below, and the performance in the lymph nodes whose cross-sectional diameter is less than 17 mm is seen in Figure 1(b):



(a) Lymph nodes of all sizes



(b) Lymph nodes with cross-section diameter <17mm

References

1. Chen L, Zhou Z, Sher D, et al. Combining many-objective radiomics and 3D convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer. *Phys Med Biol.* 2019;64(7):075011.
2. Deb K, Pratap A, Agarwal S, Meyarivan T. A fast and elitist multiobjective genetic algorithm: NSGA-II. *Evolutionary Computation, IEEE Transactions on.* 2002;6(2):182-197.
3. Wang S, Hou Y, Li Z, Dong J, Tang C. Combining convnets with hand-crafted features for action recognition based on an HMM-SVM classifier. *Multimedia Tools and Applications.* 2016:1-16.
4. Antropova N, Huynh BQ, Giger ML. A Deep Feature Fusion Methodology for Breast Cancer Diagnosis Demonstrated on Three Imaging Modality Datasets. *Medical physics.* 2017.
5. Li S, Xu P, Li B, et al. Predicting lung nodule malignancies by combining deep convolutional neural network and handcrafted features. *Physics in Medicine & Biology.* 2019;64(17):175012.
6. Chen L, Zhou Z, Sher D, et al. Combining many-objective radiomics and 3-dimensional convolutional neural network through evidential reasoning to predict lymph node metastasis in head and neck cancer. *Physics in medicine and biology.* 2019.