

SUPPLEMENTARY MATERIALS

Supplementary Methods

Nuclei segmentation

The deep learning model developed by Mahmood, et al.¹ was employed for the task of nuclear segmentation. Deep learning is a type of machine learning technique that attempts to learn by example using neural networks with multiple layers². Given a set of raw data, a deep-learning algorithm tries to discover what features are relevant. It iteratively improves upon learned representations of the underlying data with the goal of maximally attaining class separability³.

More specifically, the work of Mahmood et al. uses a generative adversarial network, a deep learning framework that learns from a set of training data and generates new data with the same characteristics as the training data. According to the authors, the use of this framework facilitates to capture higher-order statistics from images, so the resulting networks are more context-aware.

For our study, the model of Mahmood et al.¹ was not re-trained but used without modification. It receives as input a 2048x2048-pixel H&E image patch, and outputs a segmentation mask indicating which image pixels correspond to nuclei.

Lymphocyte detection

The approach developed by Corredor et al.⁴ was utilized for detecting lymphocytes. This method starts by applying image color normalization to compensate staining variations of slides acquired from different institutions. Then, a set of visual features related to texture, shape, and color are extracted from each segmented nucleus considering that lymphocytes are generally distinguished from other cell nuclei by their smaller size, more circular shape, and darker homogeneous staining. These visual features are then used to train a machine learning model (a support vector machine with linear kernel) that classifies each nucleus as either a lymphocyte or a non-lymphocyte.

For our study, this lymphocyte detection method⁴ was not re-trained or modified. The model receives as input both a 2048x2048-pixel H&E image patch and its respective nuclei segmentation mask (obtained using the method of Mahmood et al.¹), and it outputs the location of lymphocytes and non-lymphocytes within the image.

Although that lymphocyte model⁴ was trained on lung images, lymphocytes are very similar appearing across different organs. Nonetheless, to ensure this model was able to identify TILs on HPV-associated OPSCC correctly, we ran a validation experiment: An expert pathologist visually examined sixty 512x512-pixel tiles, randomly extracted from datasets D1-D6 (ten per dataset). The pathologist checked the quality of TIL detection on each tile and assigned each tile into either an excellent, good, fair, or poor category. The pathologist ranked 60% of the tiles as excellent, 18.3% good, 15% fair, and 6.7% poor. These results suggested the performance of the segmentation model was sufficiently accurate for the OP-TIL classifier.

Analysis of correlation between OP-TIL risk scores for DFS and OS

We have computed the Pearson's correlation coefficient for risk scores of DFS and OS in training. The correlation was moderate = 0.5178. Additionally, we observed that, most of the patients classified as "high risk" by the DFS model were also classified as "high risk" by the OS model (n=248). However, there were some cases (n=81) in which a patient was classified as "high risk" by the DFS model and "low risk" by the OS model. Similarly, some patients (n=16) classified as "low risk" by the DFS model were set as "high risk" by the OS model.

We hypothesize two possible reasons for this: 1) OS information could be noisy since it includes deaths from any cause, i.e., we do not necessarily have disease-specific death information; 2) A patient may have biological characteristics that make him/her more prone to recurrence than to death and vice versa.

References

1. Mahmood F, Borders D, Chen R, et al. Deep Adversarial Training for Multi-Organ Nuclei Segmentation in Histopathology Images. *IEEE Trans Med Imag*. Published online 2019.
2. LeCun Y, Bengio Y, Hinton G. Deep learning. *Nature*. 2015;521(7553):436-444. doi:10.1038/nature14539
3. Janowczyk A, Madabhushi A. Deep learning for digital pathology image analysis: A comprehensive tutorial with selected use cases. *J Pathol Inform*. 2016;7(1):29. doi:10.4103/2153-3539.186902
4. Corredor G, Wang X, Lu C, Velcheti V, Romero E, Madabhushi A. A watershed and feature-based approach for automated detection of lymphocytes on lung cancer images. In: *SPIE Medical Imaging*. International Society for Optics and Photonics; 2018.
5. Thompson LDR, Burchette R, Iganej S, Bhattasali O. Oropharyngeal Squamous Cell Carcinoma in 390 Patients: Analysis of Clinical and Histological Criteria Which Significantly Impact Outcome. *Head Neck Pathol*. 2019;14(3):666-688. doi:10.1007/s12105-019-01096-0
6. Li W, Cerise JE, Yang Y, Han H. Application of t-SNE to human genetic data. *J Bioinform Comput Biol*. 2017;15(04):1750017. doi:10.1142/S0219720017500172

Supplementary Tables

Supplementary Table 1. Top OP-TIL features for disease-free and overall survival.^a

Top OP-TIL features for survival measured at each tile	Statistic for WSI	Feature weight
Disease-free Survival		
Mode of the Compactness of TILs	Standard deviation	-5.7124
Total percentage of non-TIL clusters surrounding TIL clusters when looking at the closest cluster	Minimum	-0.2393
Mean of the percentage of non-TIL clusters surrounding other non-TIL clusters when looking at the two closest clusters	Maximum	-0.2044
Mean of the density of non-TIL clusters	Standard deviation	-0.1687
Skewness of the density of TIL clusters	Skewness	-0.1650
Minimum percentage of TIL clusters surrounding other TIL clusters when looking at the closest cluster	Skewness	0.0216
Maximum percentage of non-TIL clusters surrounding TIL clusters when looking at the closest cluster	Minimum	-4.63e-15
Overall Survival		
Ratio between the number of lymphocytes and the total area of the convex hull	Standard deviation	-303.42
Mode of the Compactness of TILs	Standard deviation	-0.3757
Skewness of $\frac{AF1 \cap AF2}{AF2}$, with <i>AF1</i> the area of a convex hull containing the centroids of all the TIL clusters and <i>AF2</i> the area of a convex hull containing the centroids of all the non-TIL clusters	Skewness	0.0565
Average edge length of the minimum spanning tree of non-TIL clusters.	Kurtosis	-0.0057
Kurtosis of the area of TIL clusters	Standard deviation	-0.0032
Skewness of compactness of TIL clusters	Kurtosis	-0.0021
Minimum value of the TIL density matrix	Standard deviation	-0.0017
Maximum area of TIL clusters	Mean	-9.41e-08

^a WSI = whole-slide image; TIL = tumor-infiltrating lymphocyte.

Supplementary Table 2. Univariable and multivariable survival analyses for overall survival including all comers (≤ 30 pack-year smoking history) in the testing sets (D2-D6).^a

Variable	Univariable		Multivariable	
	HR (95% CI)	<i>P</i> ^b	HR (95% CI)	<i>P</i> ^b
Age (≥ 55 vs. < 55 years) ^c	2.23 (1.11-4.50)	0.04	1.11 (1.07-1.16)	< 0.001
Smoking (≥ 10 vs. < 10 pack-year)	1.44 (0.70-2.98)	0.30	0.98 (0.94-1.01)	0.20
T-stage (T1 vs. T2)	2.42 (1.20-4.90)	0.02	2.44 (1.10-5.89)	0.03
N-stage (N0 vs. N1)	^d	0.15	5.84 (0.78-748.21)	0.10
Treatment (Surgery + AT vs. others)	1.63 (0.79-3.34)	0.17	1.63 (0.79-3.38)	0.18
OP-TIL (Low- vs. high-risk)	2.34 (1.08-5.07)	0.02	2.58 (1.09-5.68)	0.03

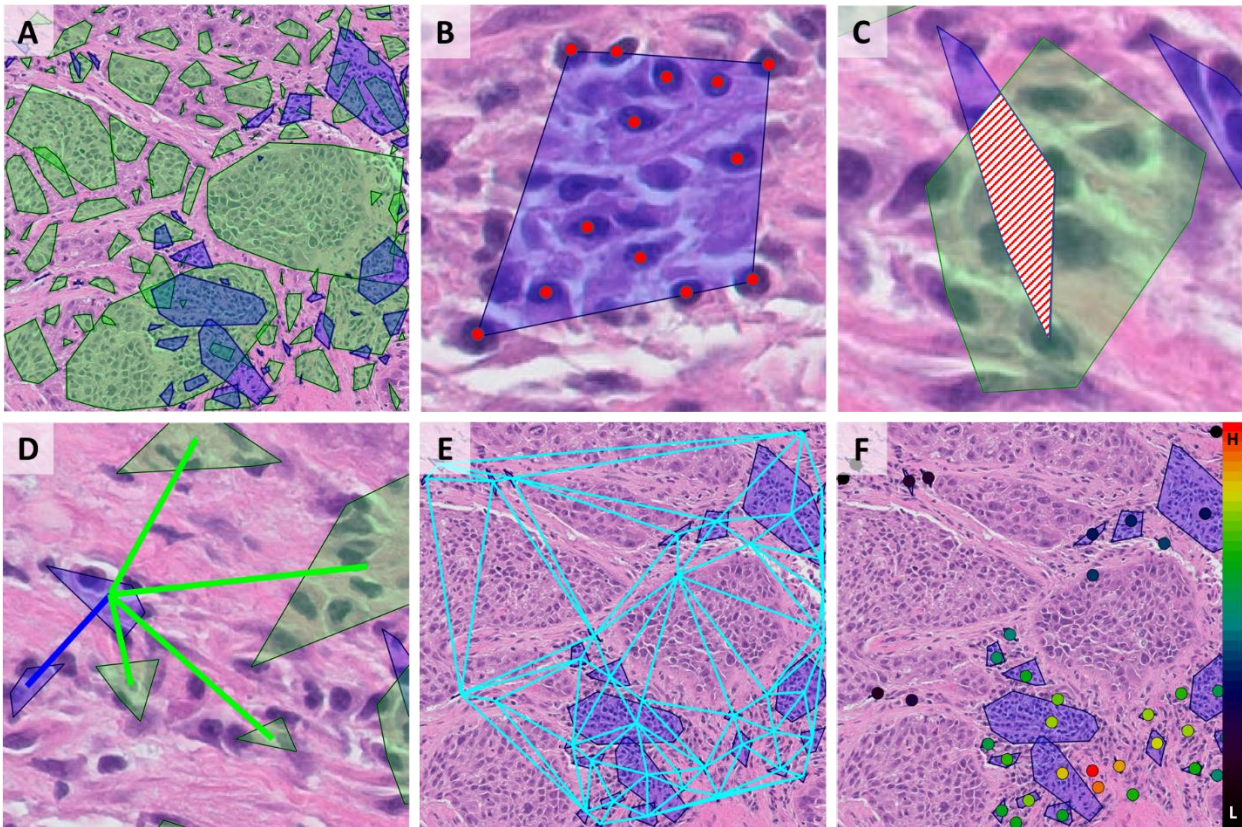
^aFor univariable analysis, age and smoking were dichotomized while for multivariable they were used continuously. HR = hazard ratio; CI = confidence interval; AT = adjuvant therapy.

^b *P* values were two-sided and computed using the log rank test.

^cThe cutoff for age was set to 55 years, as suggested by Thomson, et al.⁵

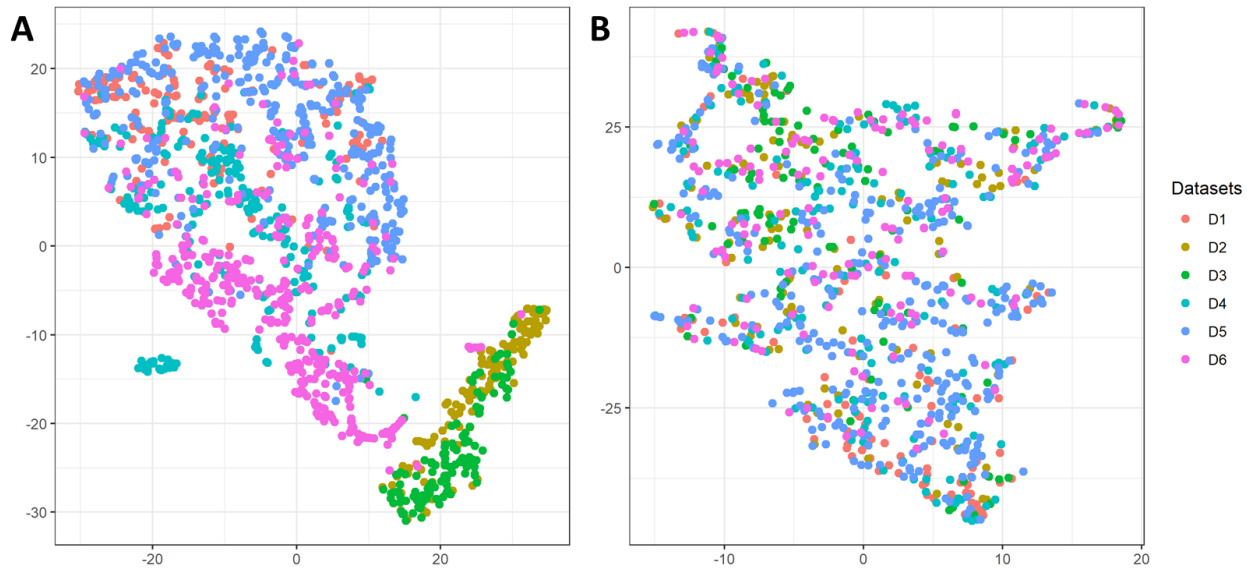
^dHazard ratio was not computed for N-stage since none of the N0 stage patients had death events.

Supplementary Figures

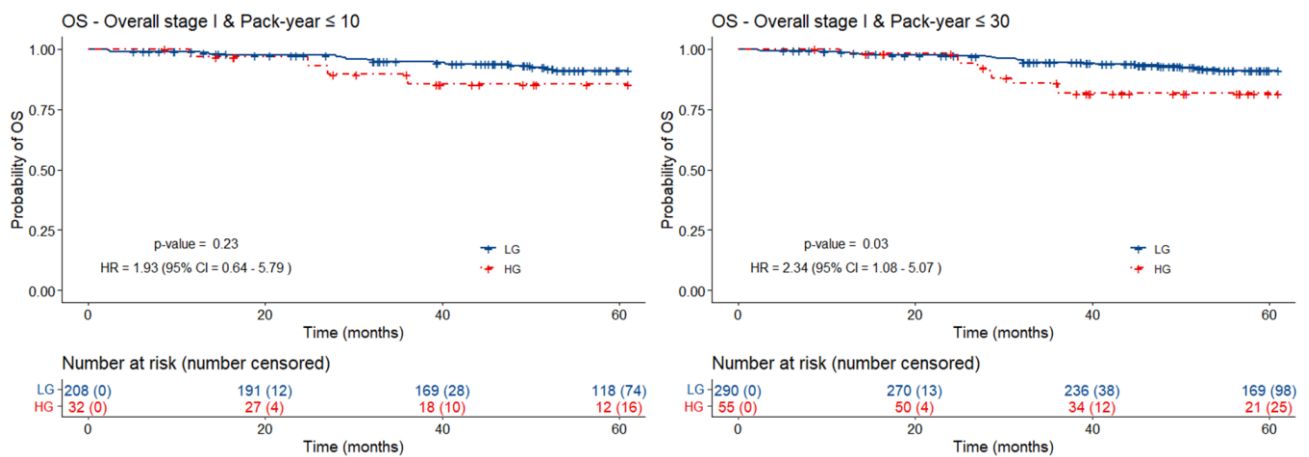


Supplementary Figure 1. Computation of some representative OP-TIL features. (A) Number of clusters of TILs (blue) and non-TILs (green) in a tile. (B) Density of TILs in a specific cluster computed as the ratio between the number of TILs (red dots) and the area of the convex hull (blue). (C) Area intersected between clusters of TILs and non-TILs. (D) Cluster neighborhood diversity. (E) Delaunay graph built for TIL clusters. (F) Node compactness of TIL clusters. The color bar represents the grouping measurement, in which H stands for nodes highly grouped, i.e., very close to multiple nodes, while L stands for lowly clustered nodes, i.e., isolated or far from other nodes.

TIL = tumor-infiltrating lymphocyte.



Supplementary Figure 2. A total of 985 p16-positive OPSCC patients from six different sites were embedded into a two-dimensional feature space and then plotted using the t-stochastic neighbor algorithm⁶. Each point represents a patient and each color a different site. The embedding was done using (A) image metrics related to brightness and contrast (computed using HistoQC) and (B) OP-TIL features. Panel A shows that patients tend to form clusters in site-specific groups due to batch effects. However, Panel B shows that there are no evident clusters of patients by institutional site (cohort). This suggests that OP-TIL features are resilient to batch effects and are reproducible across the multiple sites/cohorts. [OPSCC = oropharyngeal squamous cell carcinoma](#)



Supplementary Figure 3. Kaplan–Meier plots for the OS OP-TIL classifier applied to patients in the validation set (D2–D6) with overall stage I (AJCC 8th ed.) and with less than 30 pack-year of smoking history. Patients with less than 30 pack-year classified by OP-TIL as “high risk” (dashed line) are approximately 2 times more likely to die. P values were two-sided and computed using the log rank test. OS = overall survival; AJCC = American Joint Committee on Cancer; HR = hazard ratio; CI = confidence interval; LG = low-risk group; HG = high-risk group.