Modular validation of lymphocyte detection and tumor and stroma segmentation models to accurately predict tumor-infiltrating lymphocytes from H&E images in metastatic lung adenocarcinoma

Bolesław Osinski[1], Qiyuan Hu[1], Sun Hae Hong[1], Kunal Nagpal[1], Ben Terdich[1], Yoni Muller[1], Arlen Brickman[1], Nike Beaubier[1]
[1]Tempus AI, Inc., Chicago, IL
Disclosure Statement: All authors are employees of Tempus AI, Inc. and hold stock in the company.

INTRODUCTION

- Accurate quantification of tumor-infiltrating lymphocytes (TIL) can predict patient response to immune checkpoint inhibitor (ICI) therapy in lung adenocarcinoma (LUAD)
- We developed an AI model, consisting of 2 models: 1) lymphocyte detection, 2) tumor & stromal ("scorable") region segmentation), to predict the density of lymphocytes within LUAD tumors
- Manual whole-slide TIL scoring methods are subjective and inconsistent, so we validated each model as an LDT, with high quality ground truth, using IHC-derived labels for lymphocytes and a consensus of 4 pathologists for regions.

DESIGN

Lymphocyte model: 280 slides were stained for H&E, scanned, de-stained with xylene and re-stained with CD3/CD20 (T-Cell/B-Cell lymphocytes), and scanned again. In QuPath, pathologists annotated 5-10 fields of view (FOVs, 64x64μm^2) per slide and initiated stain detection to label each cell as Lymphocyte or Other. These were registered to the corresponding H&E, where pathologists edited labels, & errors. Slides with poor IHC staining or failed registration were removed, & slides were split into train (110) / tuning (56) / test (62) sets. We used a UNet with customized cross-entropy loss function on the point labels. To evaluate predictions, points within 3μm of each other were assumed to label the same cell.

Tumor and stroma region model: Tumor and stroma region annotations were performed directly on H&E slides (N=266) within FOVs (1mm^2) by 4 pathologists per FOV. Data was split into train (140) / tuning (60) / test (66) sets, and a UNet model was trained. Following an initial evaluation, the training set was supplemented with annotations on metastatic breast cancer (N=308), found to improved performance. The model predictions are post-processed to obtain a "scorable region", defined as the tumor region and 50μm of tumor-associated stroma.

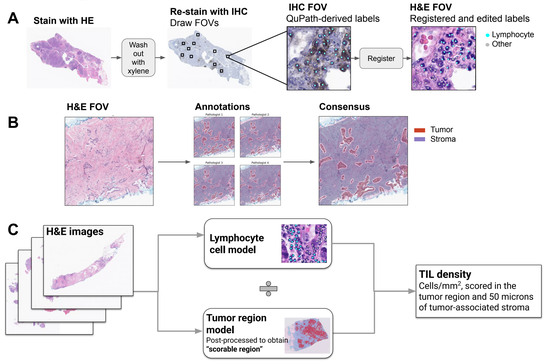


Figure 1. Ground truth collection for lymphocyte (A) and region (B) models. (C) Schematic of TIL density computation.

SUMMARY

- Both models meet the analytical validation acceptance criteria for our Laboratory-Developed Test (LDT).
- The region model demonstrates strong performance across all characteristics, while the lymphocyte detection model performs well in primary lung and lymph nodes but could be improved in other metastatic sites.
- Both models are robust to scanner variations and artifacts

RESULTS

Table 1. Summary of Analytical Validation of Both Models

	Cohort Sizes		Analytical Validation Results	
Characteristic	**Lymphocyte N slides (N FOVs)**	**Scorable region N slides (N FOVs)**	**Lymphocyte CCC***	**Scorable Region F1***
Tissue Site				
Lung	33 (313)	32 (63)	0.93	0.87
Lymph node	13 (119)	13 (24)	0.92	0.87
Liver	3 (25)	5 (10)	0.38	0.91
Bone	4 (35)	5 (10)	0.17	0.92
Adrenal gland	5 (47)	6 (12)	0.59	0.85
Soft tissue	4 (40)	5 (9)	0.81	0.88
Procedure Type				
Resection & Excision	15 (150)	11 (22)	0.95	0.91
Core needle biopsy	47 (429)	55 (107)	0.87	0.86
Subtype (lung tissue)				
Acinar	10 (88)	10 (20)	0.87	0.88
Mucinous	9 (90)	8 (16)	0.94	0.86
Solid	6 (60)	5 (10)	0.83	0.84
Lepidic	2 (15)	3 (6)	0.92	0.74
Papillary	2 (20)	0 (0)	0.97	N/A
Micropapillary	2 (20)	2 (4)	0.97	0.96
Unspecified	2 (20)	4 (7)	0.93	0.90
Total (all tissues)	62 (579)	66 (128)	0.90	0.87

Acceptance criteria: CCC > 0.85 F1 > 0.70

*CCC: concordance correlation coefficient
*F1 (DICE score) is computed from a single bulk confusion matrix which is summed from all FOVs.

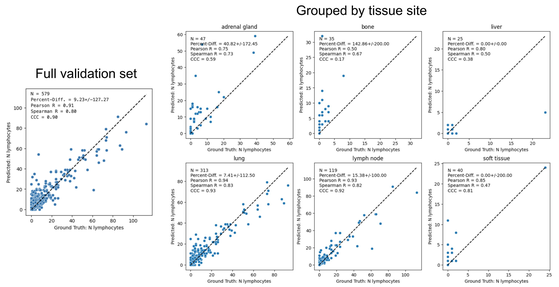


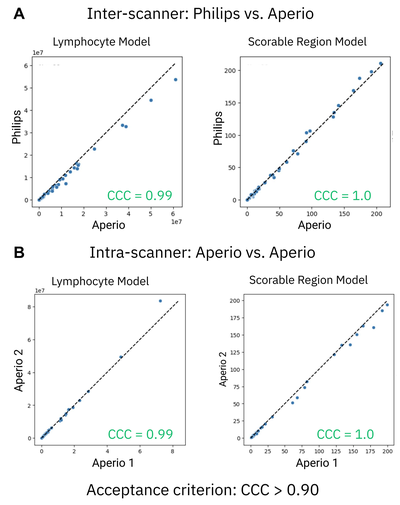
Figure 2. Analytical Accuracy: Lymphocyte model detail. Scatter plots comparing lymphocyte model predictions to ground truth, with each point representing the cell count from a field of view (FOV). Bone, liver, and soft tissue exhibit extremely low cell counts, potentially making them unsuitable for the CCC metric.

Table 2. Limit of Detection (LoD) Simulation

N FOVs	Area (mm^2)	Mean Pearson R	Mean RMSE	Mean CCC
3	0.01229	0.628	880.573	0.565
5	0.02048	0.691	740.449	0.641
10	0.04096	0.802	503.425	0.789
15	**0.06144**	**0.851**	**415.014**	**0.845**
20	0.08192	0.859	397.514	0.853
30	0.12288	0.907	312.584	0.906
50	0.20480	0.938	252.723	0.937

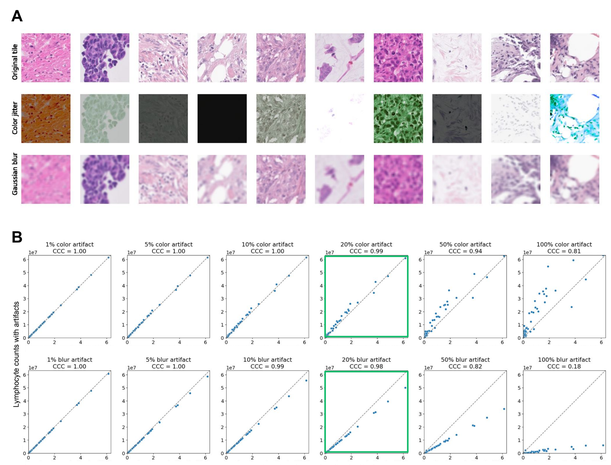
Simulation method

Successively fewer FOVs of size (64x64) μm^2 (50 down to 3), were randomly sampled 10 times from the predicted scorable region, ensuring FOVs are 100% filled by scorable region. Correlation between TIL densities of sampled regions and the whole slide was measured. The LoD is set at 15 FOVs (0.06mm^2), where mean CCC > 0.8.



Acceptance criterion: CCC > 0.90

Figure 3. Analytical Precision: Scanner variability. Model predictions demonstrate strong robustness across scans from Leica Aperio GT450 and Philips UFS scanners (A) and between rescans on Leica Aperio GT450 scanners (B).



Acceptance criteria: CCC > 0.95 **Result:** maximum allowable artifact percentage Color artifact: 20% | Blur artifact: 20%

Acceptance criteria: F1 > 0.80 **Result:** maximum allowable artifact percentage Color artifact: 20% | Blur artifact: 50%

Figure 4. Analytical Specificity: Robustness to Artifacts. (A) Example tiles illustrating the effects of color transformations and Gaussian blurring used to simulate artifacts. (B) Simulation results for the lymphocyte model, with color artifacts (top) and blur artifacts (bottom). (C) Simulation results for the scorable region model, with color artifacts (left) and blur artifacts (right). Both models exhibit strong robustness to these artifacts.

ACKNOWLEDGMENTS

We thank Dana DeSantis from the Tempus Science Communications team for poster development.

Correspondence: bo.osinski@tempus.com