Institute for Data Science in Oncology

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center

PS3-04-26

Reproducibility and Robustness of an AI-based Stromal Tumor-Infiltrating Lymphocyte Pipeline in Triple-Negative Breast Cancer

Sara Ranjbar[1,2,*], Anthony San Lucas[3,*], Pingjun Chen[1,2], Khushali Shah[3], Michael Suchko[3], Bella Guerrouahen[3], Caner Ercan[1], Elizve Barrientos Toro[1], Keith Sweeney[4], Xiaoxi Pan[1,2], Jessica DeLa Cruz[3], B. Leticia Rodriguez[1,2], Neelima Reddy[3], Clinton Yam[5], Geeta Mantha[3], Michael Riben[4], Lei Huo[4,^], Yinyin Yuan[1,2,^]

[1]Department of Translational Molecular Pathology, [2]Institute for Data Science in Oncology, [3]Division of Pathology and Lab Medicine Clinical Trial Operations, [4]Department of Pathology, [5]Department of Breast Medical Oncology, The University of Texas MD Anderson Cancer Center, Houston, TX,USA.

*contributed equally. ^share senior authorship. Contact Info: pchen6@mdanderson.org

THE UNIVERSITY OF TEXAS MD Anderson Cancer Center Making Cancer History®

LYDA HILL PHILANTHROPIES

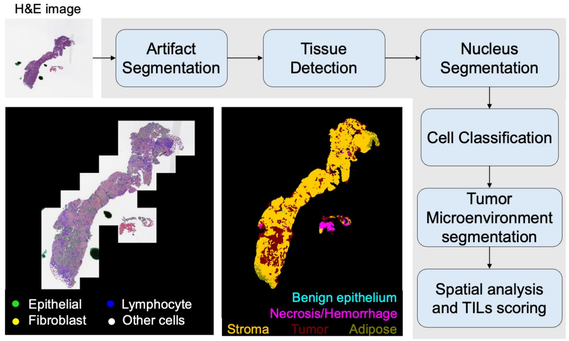
Background

Triple-negative breast cancer (TNBC) has limited treatment options and severe treatment-related toxicity. Recent studies have shown that quantifying stromal tumor-infiltrating lymphocytes (sTILs) can aid in identifying early-stage TNBC patients likely to achieve a pathological complete response (pCR) to neoadjuvant chemotherapy [1], suggesting its utility in treatment stratification. However, manual assessment of sTILs is prone to interobserver variability. Artificial intelligence (AI) promises a more consistent alternative, but reproducibility and robustness of AI to input variability remains insufficiently evaluated, limiting clinical adoption of pathology AI tools [2].

Methods

We used our in-house AI-sTIL pipeline [3] (Figure 1), developed using H&E-stained slides from core biopsy samples in the ARTEMIS trial (NCT02276443), scanned between 2016 and 2024 with a Leica Aperio AT2 scanner at 20x magnification. We compared AI-generated scores (ai_stil) with two independent manual sTILs provided by two pathologists (n=201; 42% pCR+) according to the International TIL Working Group guidelines.

We evaluated reproducibility and robustness of AI-sTIL to operational variations by generating new images from 150 ARTEMIS patients. We scanned using different scanners and magnification levels, changed operators, repeated scanning and performed pairwise comparisons across controlled groups to identify potential biases. In total 1,147 new scans were generated (Figure 2).



$$\text{ai_stil score} = \frac{\text{Number of Lymphocytes}}{\text{Area of stroma}}$$

Figure 1. AI-sTIL pipeline.

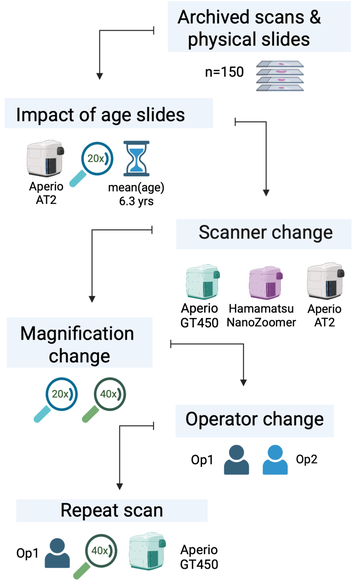


Figure 2. Data collection scheme.

Results

- We observed high sTILs agreement between pathologists (ρ=0.884, R^2=0.855; Figure 3a). The AI score correlated well with individual manual sTILs from Pathologist #1 (ρ=0.768, R^2=0.502; Figure 3b) and Pathologist #2 (ρ=0.766, R^2=0.579; Figure 3c), and showed stronger correlation with the average of the two (ρ=0.797, R^2=0.564; Figure 3d), underscoring potential to serve as a consensus.

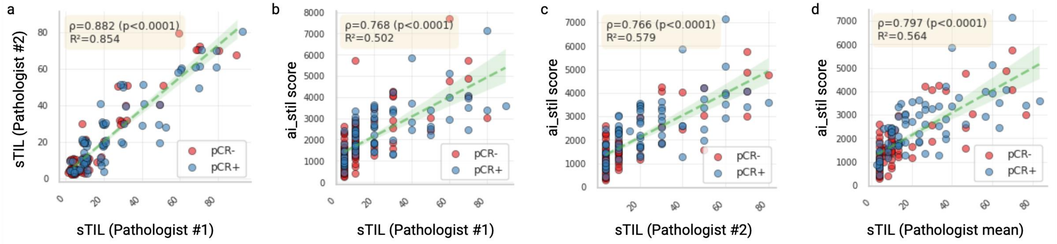


Figure 3. Manual sTILs vs. AI-sTIL. (a) Inter-pathologist agreement; (b) AI vs. Pathologist #1; (c) AI vs. Pathologist #2; (d) AI scores vs. average of the two sets of sTILs.

- AI scores demonstrated high reproducibility across all tested variations (Figure 4; Spearman ρ>0.972, R^2>0.941, p<0.001 for all comparisons). A minor bias was observed with scanner changes, particularly when combined with magnification change, indicating that calibration or standardized digitization protocols may be required.

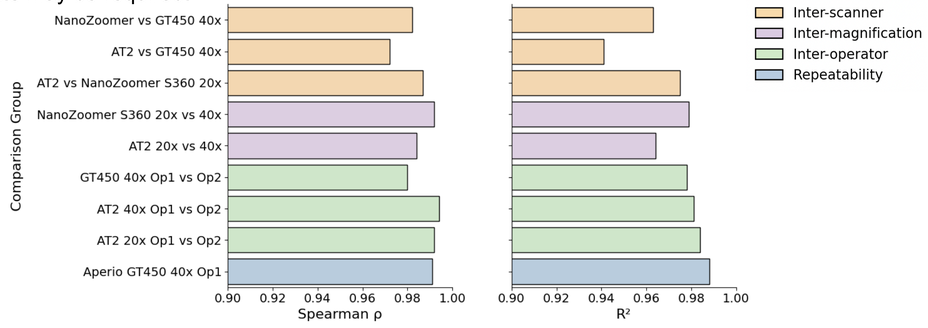


Figure 4. Reproducibility and robustness evaluation outcomes.

- We observed a trend towards lower scores for older slides (ρ=-0.366; 5.6% decrease in score per year, Figure 5), likely attributable to stain fading. However, such prolonged delays between tissue sectioning and scanning are uncommon in routine practice. Nevertheless, we incorporated a 2-month time limit between slide preparation and scanning into the use instructions of the pipeline.

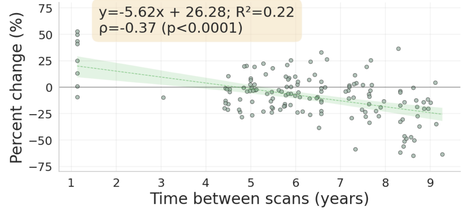


Figure 5. Impact of slide age in AI-sTIL scores.

Conclusion

Our work highlights the reproducibility and robustness of the AI-sTIL pipeline across common variations in imaging data and the importance of identifying safe boundaries of AI use. Our approach provides a practical framework for institutions seeking to assess deployment readiness of pathology AI tools before formal validation and clinical integration.

References

1. Abuhadra *et al.*, Cancers (2023).
2. Park *et al.*, JAMIA Open (2020).
3. Pan *et al.*, Clinical Cancer Research (2025).

Acknowledgements

This work is supported by the Spatial Ecology & QUantitative pathOlogy Image Analytical platform (SEQUOIA) through the MD Anderson STrategic Research Initiative Development Program (STRIDE) at The University of Texas MD Anderson Cancer Center. This project is funded by Lyda Hill Philanthropies.