Interdisciplinary Research in Medical Sciences Specialty
ISSN: 3062-4401
2026, Volume 6, Issue 1, Page No: 271-283

Available online at: www.galaxypub.co/page/journals

Galaxy Publication

Assessing the Analytical and Clinical Validity of AI-Based Tumor-Infiltrating Lymphocyte Scoring in TNBC and the Interchangeability of Machine Learning Models

Noah Williams[1]*, Grace Collins[1], Ethan Brooks[2], Daniel Green[3]

[1]Department of Triple-Negative Breast Cancer and TIL Scoring, Faculty of Medicine, University of Sydney, Sydney, Australia.
[2]Department of AI-Based Digital Pathology, Faculty of Medicine, University of Queensland, Brisbane, Australia.
[3]Department of Machine Learning Model Interchangeability, Faculty of Medicine, University of New South Wales, Sydney, Australia.

***E-mail** ✉ noah.williams@sydney.edu.au

Received: 07 April 2026; **Revised:** 11 June 2026; **Accepted:** 17 June 2026

ABSTRACT

Manual evaluation of tumor-infiltrating lymphocytes (TILs) by pathologists has revealed important predictive and prognostic roles in both early-stage and advanced triple-negative breast cancer (TNBC). Despite this, the approach continues to suffer from inconsistency. Artificial intelligence (AI) represents an effective way to minimize such inconsistency and support fully automated, impartial TILs measurement. The main hurdle to widespread clinical use is proving strong analytical and prognostic reliability. Ten different AI models were examined for their influence on TILs quantification, highlighting variations in analytical and prognostic performance. This included seven newly created models and three already validated ones. Testing occurred in a retrospective group for analytical aspects and a separate prospective group for prognostic aspects, using invasive disease-free survival (IDFS) as the outcome measure over a median 4-year follow-up period. Model development and analytical testing drew on diagnostic slides from 79 women treated for primary invasive TNBC at Yale School of Medicine between 2012 and 2016. Prognostic evaluation used an external group of 215 TNBC cases diagnosed in Sweden from 2010 to 2015. Marked differences emerged in analytical validity among the models and their training methods (Spearman's $r = 0.63–0.73$, $p < 0.001$). Eight of the ten AI systems nonetheless displayed significant prognostic value for digital TILs scores in the independent external cohort. Hazard ratios were closely aligned and overlapping (HR = 0.40–0.47; $p < 0.004$) via Cox regression on the IDFS endpoint, even for models with more limited training. The consistent prognostic strength seen in most AI TIL models reflects the fundamental reliability of host anti-tumor immune response (captured through TILs) as a biomarker. Model-to-model differences, however, deserve attention. A shared, extensive, multi-institutional dataset is essential as a reference standard to enable fair comparisons and confirm dependability of diverse AI solutions ahead of routine clinical deployment.

Keywords: TILs, Tumor infiltrating lymphocytes, Breast cancer, Machine learning, Deep learning, Artificial intelligence

How to Cite This Article: Williams N, Collins G, Brooks E, Green D. Assessing the Analytical and Clinical Validity of AI-Based Tumor-Infiltrating Lymphocyte Scoring in TNBC and the Interchangeability of Machine Learning Models. Interdiscip Res Med Sci Spec. 2026;6(1):271-83. https://doi.org/10.51847/eopncqTqQW

Introduction

Over the last five years, major progress has occurred in developing new treatments for early breast cancer [1, 2]. Better tools are urgently needed for biomarker-based patient risk assessment to avoid unnecessary therapy or missed opportunities, allowing precise selection of individuals likely to gain from extra treatment. The amount of lymphocytes present in the tumor stroma (stromal TILs, or sTILs) visible on standard H&E-stained FFPE slides stands out as a powerful prognostic marker in early triple-negative breast cancer (TNBC) [3–8], where elevated sTILs can meaningfully lower risk estimates beyond traditional clinicopathological staging [9]. The International

Immuno-Oncology Biomarker Working Group on Breast Cancer (TIL-WG) has issued recommendations to promote uniform sTILs evaluation with proven consistency [5, 10–12]. Even so, variability between observers cannot be fully eliminated, as sTILs assessment is inherently semi-quantitative and struggles to represent the full complexity of the tumor-immune microenvironment (TME).

Digital image analysis powered by machine learning can overcome standardization problems and detect subtle image patterns invisible to human observers. This may enable richer, more dependable outcome predictions than straightforward TILs counts [13–16].

Confirming adequate analytical and clinical validity is the critical barrier to bringing AI-TILs tools into everyday use. Without this, applying them for treatment decisions carries unacceptable risk [17–19]. Large, high-quality datasets for training, testing, and validation are vital to stop creation of overfitted models that succeed only on similar data and fail externally [20, 21]. Evidence comparing how various AI systems perform on TILs evaluation is also inadequate. This study therefore evaluates the analytical and prognostic characteristics of ten AI TIL assessment tools, benchmarking them against expert manual sTILs scoring by two pathologists.

Materials and Methods

Patient cohorts

Yale cohort: analytical validity

Whole tissue section slides from 106 women diagnosed with primary invasive (stage I-III) TNBC between 2012 and 2016 were collected from the archives of the Yale School of Medicine Department of Pathology. Retrieval occurred under Yale Human Investigation Committee protocol #9505008219 to DLR [22]. One slide per case was chosen for four patients who had duplicates, based on quality regarding artifacts and presence of invasive cancer. Thirteen slides were removed due to missing invasive tumor or major artifacts. In total, 92 slides belonging to 79 patients **(Table 1)** entered the study for model training (N = 50) and internal analytical validation (N = 42). Allocation to training and test sets was performed at the patient level to eliminate any information leakage.

Table 1. Clinical characteristics of patients included in the study.

	Yale cohort	Swedish cohort	p-value[b]
Cases	79[a]	215	–
Age			
<50	21 (26.6%)	49 (23%)	0.6
≥50	58 (73.4%)	166 (77%)	
Tumor size			
≤20	37 (46.8%)	108 (50.2%)	0.7
>20	42 (53.2%)	107 (49.8%)	
Continuous (mm)	21 (15–30.5)	21 (15–30)	0.78
Histologic grade			
2	17 (21.5%)	22 (10.2%)	0.02
3	61 (77.2%)	190 (88.4%)	
NA	1 (1.3%)	3 (1.4%)	
Nodal status			
Negative	52 (65.8%)	132 (61.4%)	0.64
Positive	27 (34.2%)	81 (37.7%)	
NA	–	2 (0.9%)	
Chemotherapy			
No	12 (15.2%)	57 (26.5%)	0.46
Yes	46 (58.2%)	158 (73.5%)	
NA	21 (26.6%)	–	
Follow-up (years)			
Follow-up time	10.63 (6.3–16.4)	3.94 (3.2–5.32)	–

For dichotomous variables, number of samples and prevalence percentages are reported.

For continuous variables, median as well as first and third quartiles are reported.

a. EQC cohort: 92 slides from 79 patient cases.

b. Chi-squared test for categorical variables and Mann–Whitney test for continuous variables.

Swedish cohort: clinical validity

To independently confirm the prognostic capabilities of every AI model, researchers utilized an external Swedish group that included whole tissue sections along with detailed clinical records from 215 women who had surgery for TNBC. Participants were drawn from the SCAN-B study, a forward-looking, observational, population-wide investigation (ClinicalTrials.gov ID NCT02306096) [23, 24] spanning diagnoses made from 2010 to 2015, with prior descriptions available [25] **(Table 1)**. Ethical oversight for the SCAN-B project came from the Regional Ethical Review Board in Lund, Sweden, under approval numbers 2009/658, 2015/277, 2016/742, 2018/267, and 2019/01252. Due to difficulties in image handling, three individuals were left out of the final evaluation. Since triple-negative breast cancer makes up roughly 10–15% of all breast malignancies [26], the assembled groups—above all the external testing set containing 215 individuals—adequately mirror the general TNBC patient population. Across the cohorts, the single characteristic with a meaningful statistical distinction was histological grade, an aspect recognized for considerable disagreement among pathologists [27] **(Table 1)**.

Digital-image analysis

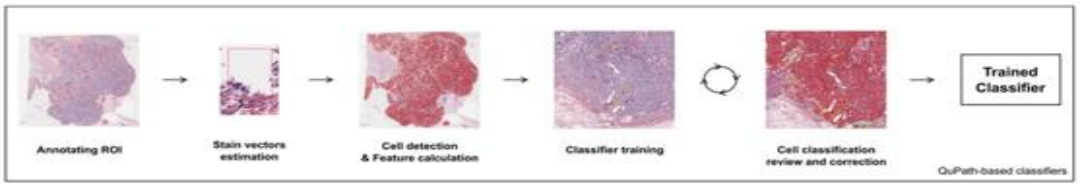
Image data acquisition and processing techniques

Slides stained with H&E were digitally scanned in full at 20× objective using the Leica Aperio ScanScope equipment running Controller version 10.2.0.2359 and ScanScope Console version 10.2.0.2352 [22]. In the case of the Swedish cohort's whole tissue sections, scanning took place on a NanoZoomer 2.0-HT device (Hamamatsu Photonics K.K.) at 20×, yielding pixels measuring 0.4537 × 0.4537 mm [25].

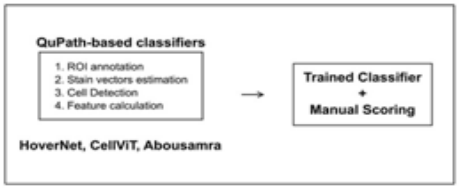
Supervised machine learning models

Development of the automatic TILs assessment tools relied on the freely available QuPath platform (release 0.3.2) [28]. The overall process for analyzing tissue images quantitatively is outlined in **Figure 1**. First, experts outlined regions of interest (ROI) that contained only invasive carcinoma following the standards set by the International TIL Working Group (ITWG) [11]. Specific rules applied were: (i) Include lymphocytes situated within the invasive tumor's limits, spanning both central areas and the advancing edge. (ii) Register every mononuclear inflammatory cell such as lymphocytes and plasma cells, while leaving out granulocytes. (iii) Disregard any TILs positioned outside the tumor's perimeter. (iv) Leave out lymphocytes near ductal carcinoma in situ components or healthy glandular structures. (v) Bypass zones impacted by compression damage, necrotic tissue, or scarring with hyalinization. Color deconvolution via stain-vector calculation was done separately per slide to standardize the appearance of different stain elements. Cell separation employed the watershed algorithm with standard configuration values. To enrich cell-level data, researchers derived smoothed features from objects within 25 μm and 50 μm neighborhoods.

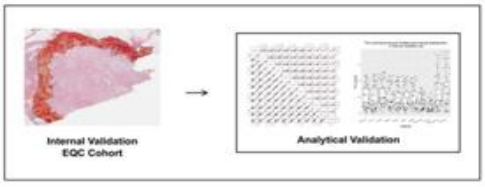
Three categories of classifiers—K Nearest Neighbor (KNN), Random Trees (RT), and Neural Network (NN)—received initial training on ten sample images. After reviewing results, the top model underwent additional rounds of training with expanded datasets, specifically incorporating 20, 30, 40, and 50 patient cases. Model designations follow the format MN, in which M stands for the algorithm family and N reflects the count of training cases.



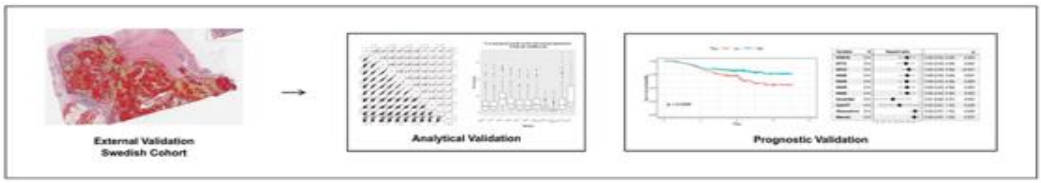
a)



b)



c)



d)

Figure 1 Overview of the digital image analysis workflow used for building and deploying the classifiers.

(a) Preprocessing steps and training pipeline for the classifiers (KNN10, RT10, NN10, NN20, NN30, NN40, and NN50). (b) Deployment of the trained TILs models. (c) Assessment of analytical performance on the Yale internal validation group. (d) Evaluation of prognostic performance in a separate external validation group. Note that the "trained classifier" shown in panels b–d corresponds to the one developed in panel a, along with the independent models HoverNet, CellViT, and Abousamra's.

Training for every classifier followed a human-in-the-loop approach that involved repeated cycles of inspection and refinement until satisfactory results were reached. Initially, manual labels were assigned to roughly 450 cells per image, ensuring at least 150 tumor cells and 150 immune cells (lymphocytes). The remaining approximately 150 cells were labeled as stroma or "other" categories (for instance, necrosis or epithelial cells). These annotations were performed across the full invasive tumor region to capture sample heterogeneity. Separate annotation sets were created for each classifier to prevent bias carry-over from prior interactive training sessions. This strategy avoided potential influence if overlapping regions of interest (ROIs) had been shared between successively trained models. Further annotations were incorporated whenever classification accuracy fell short. The performance threshold was defined as 95% accuracy inside QuPath's built-in classifier training interface. This annotation and training process was carried out by JMV under the guidance of experienced pathologists BA and JH.

Deep-learning models

Three deep-learning approaches served as standalone predictors in the analysis: HoverNet [29], CellViT [30], and the model by Abousamra *et al*. [31]. The first two are cell segmentation convolutional neural networks (CNNs) designed for cell identification and classification. Both were pre-trained on the PanNuke dataset [32], which includes nearly 200,000 nuclei from 19 different tissue types, among them breast tissue. Their output categories—neoplastic epithelial, inflammatory, connective, necrotic, and non-neoplastic epithelial cells—were mapped to tumor, immune, stroma, and other classes for subsequent TILs calculation in QuPath. In contrast, Abousamra's model was developed across 23 cancer types for patch-based classification, determining whether small image tiles (50×50 μm²) contained lymphocytes rather than performing individual cell detection [31]. Despite the clear methodological differences from the other tools, including this model as an independent reference point was considered useful for the overall comparison.

TILs scoring

For all models relying on cell detection (everything except Abousamra's), digital TILs percentages were computed using the easTILs equation [14]:

easTILs = $100 \times$ [sum of TILs area (mm²) / stroma area (mm²)],

where stroma area (mm²) = total analyzed tumor region area (mm²) − sum of tumor cell area (mm²).

For Abousamra's model, TILs scoring was defined as the proportion of the invasive cancer area classified as lymphocyte-containing patches. Pathologist evaluation of stromal TILs (sTILs) was performed on a non-overlapping portion of the slides, with BA reviewing one half and JH the other half, in accordance with established international TILs scoring recommendations.

Statistical analysis

Analytical model performance was measured using classification metrics derived from 2,500 annotated cells distributed across five slides. Spearman's rank correlation coefficient assessed the relationship between pathologist manual sTILs scores and the easTILs values generated by each AI model on both the internal and external validation sets. Invasive disease-free survival (IDFS) followed the STEEP criteria [33] and was calculated as the interval from diagnosis until death from any cause or any invasive breast cancer event (including local-regional or distant relapse). For Kaplan–Meier survival curves, TILs scores were binarized at a 10% threshold

[34–38] and are available in supplementary materials; survival differences were evaluated with the log-rank test. In the primary analysis, however, TILs scores were treated as continuous variables in Cox regression models because of variations in their distributions **(Figures 2 and 3)**.

Both univariate and multivariate Cox proportional hazards regressions were conducted to determine the prognostic impact of TILs scores after accounting for major clinicopathological factors: age, tumor size, histologic grade, and lymph node status. Age and tumor size were binarized using cut-offs of 50 years and 20 mm. Results from the Cox models are displayed in forest plots showing hazard ratios (HR) with 95% confidence intervals (CI) and Wald test p-values. Forest plots for binarized TILs scores appear in supplementary materials. Separate forest plots for the subgroup of patients who received chemotherapy (N = 158, representing 73.5% of the external validation cohort) are also included as supplementary files. To keep the main figures concise, hazard ratios for the adjustment variables were not shown in the multivariate forest plots; the study's central interest lies in the TILs effect. All adjustment covariates are well-established prognostic factors in TNBC. The assumption of proportional hazards was verified using the cox.zph function from R's survival package (version 3.5.8) and held true for all machine-derived TILs variables. A significance level of 5% was applied to all statistical tests. Model evaluation and statistical computations were executed in Python version 3.9 and R version 4.3.2.

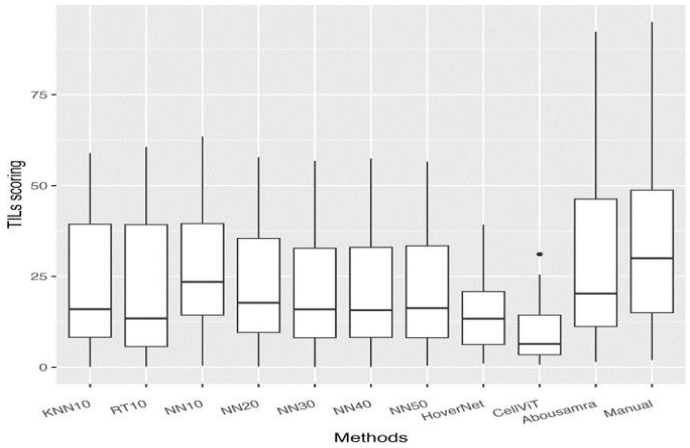


Figure 2. Boxplots displaying TILs scores produced by every method within the Yale internal validation group.

In each boxplot, the central black horizontal line marks the median value, the solid outlined rectangle spans the interquartile range (25th to 75th percentile), the thin black vertical lines extend to the overall data range, and individual dots indicate outlier values lying outside this distribution.

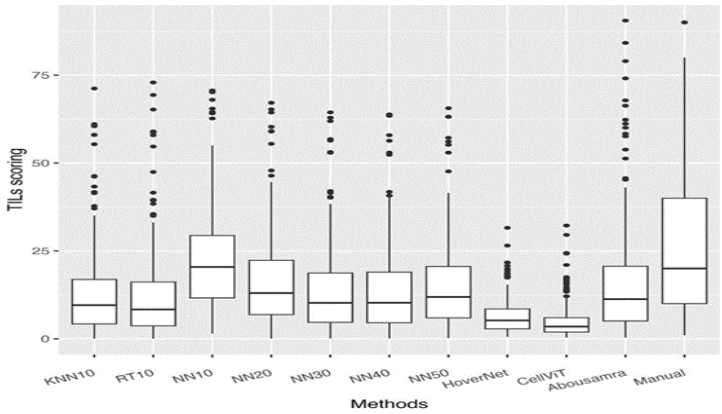


Figure 3. Boxplots illustrating TILs scores from all scoring approaches in the external SCAN-B validation group.

Within each boxplot, the thick black horizontal line denotes the median, the solid outlined rectangle covers the 25th to 75th percentile range, the thin black vertical lines indicate the full spread of the data, and scattered dots represent values that fall outside the main distribution as outliers.

Role of funding
The study sponsor had no involvement in study design and execution; data gathering, handling, examination, or interpretation; manuscript drafting, revision, or final approval; or the choice to publish the article.

Results and Discussion

Analytical validity
Evaluation of model analytical performance began with results from the Yale cohort's internal validation set. The KNN10 and RT10 approaches produced the broadest spread of TILs scores, whereas the various NN-based classifiers displayed closely aligned and stable score distributions **(Figure 2)**. In comparison, measurements derived from HoverNet and CellViT showed the most restricted distributions, while Abousamra's model generated the widest range, resembling the distribution seen in manual sTILs scoring **(Figure 2)**.
Associations between AI-generated TILs scores and pathologist manual sTILs evaluations differed notably among the models **(Figure 4)**. KNN10 demonstrated a moderate relationship with manual scores ($r = 0.72$, $p < 0.001$). NN10 performed marginally stronger ($r = 0.79$, $p < 0.001$). Among models trained on smaller sample sizes, RT10 reached the highest agreement with the reference standard ($r = 0.81$, $p < 0.001$). Expanding the training data led to progressive improvements in correlation with expert pathologist sTILs scores (NN20: $r = 0.81$, $p < 0.001$; NN30: $r = 0.82$, $p < 0.001$; NN40: $r = 0.84$, $p < 0.001$; NN50: $r = 0.83$, $p < 0.001$). HoverNet and CellViT attained the second-best correlations ($r = 0.83$, $p < 0.001$). Abousamra's model also yielded strong agreement ($r = 0.82$, $p < 0.001$).

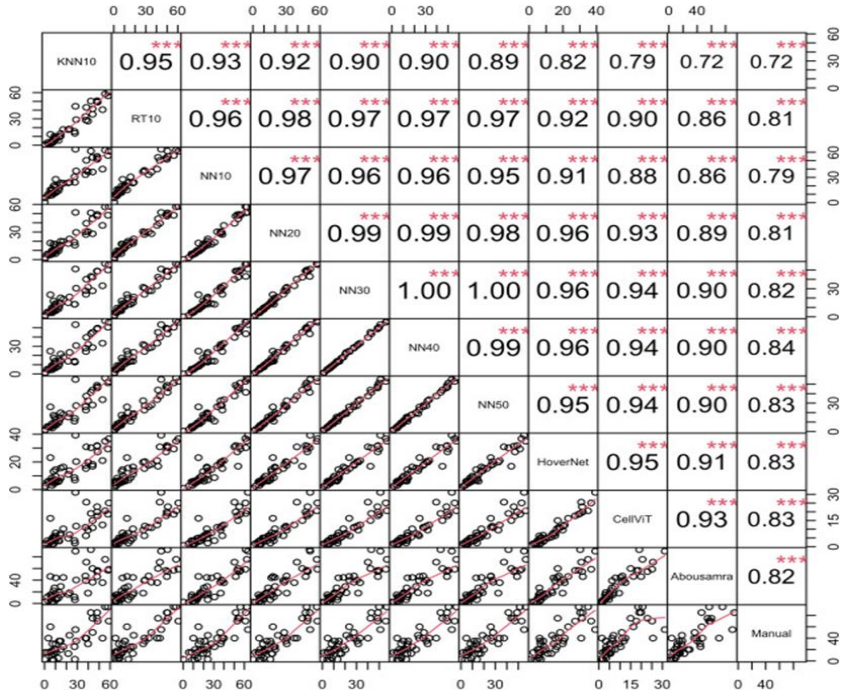


Figure 4. Matrix of Spearman correlation coefficients between every TILs evaluation approach and manual sTILs scoring in the Yale internal validation set.

The lower half displays paired scatter plots with regression lines overlaid. The upper half reports the exact correlation values and their statistical significance using asterisks, with three asterisks indicating $p < 0.001$.For a balanced and objective comparison of analytical performance across all TILs scoring techniques, the corresponding boxplot summary and correlation matrix derived from the external SCAN-B validation cohort are shown as well **(Figures 3 and 5)**. TILs score distributions proved noticeably more compact in the external group, and overall median levels were lower than those recorded internally **(Figure 3)**. Correlation strengths declined across the board. KNN10 maintained the weakest link to manual sTILs ($r = 0.63$, $p < 0.001$). Top performers were NN50 and RT10 (NN50: $r = 0.73$, $p < 0.001$; RT10: $r = 0.72$, $p < 0.001$). Interestingly, adding more training cases did not produce steady gains in alignment with expert pathologist sTILs readings (NN10: $r = 0.70$, $p < 0.001$; NN20: $r = 0.68$, $p < 0.001$; NN30: $r = 0.70$, $p < 0.001$; NN40: $r = 0.68$, $p < 0.001$; NN50: $r = 0.73$, $p < 0.001$). The

three deep-learning systems—CellViT, HoverNet, and Abousamra's—delivered correlation levels comparable to the traditional KNN, NN, and RT classifiers when benchmarked against manual scores (CellViT: $r = 0.64$, $p < 0.001$; HoverNet: $r = 0.67$, $p < 0.001$; Abousamra's: $r = 0.70$, $p < 0.001$).

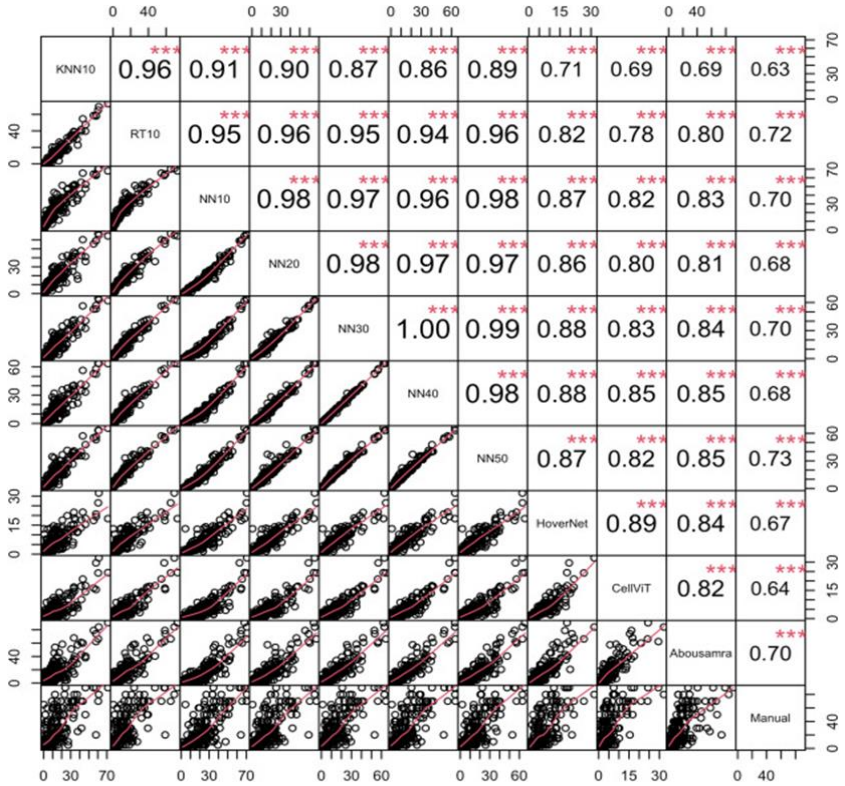


Figure 5. Spearman correlation matrix displaying relationships among all TILs evaluation methods and manual sTILs scores from the external SCAN-B validation set.

Lower triangle features paired scatter plots complete with regression lines. Upper triangle lists the precise correlation coefficients and marks significance with stars, where three stars signify $p < 0.001$.

Clinical validity

Links between the AI models and patient prognosis were analyzed within the independent Swedish SCAN-B cohort, employing invasive disease-free survival (IDFS) as the key outcome measure. Survival curves generated by the Kaplan–Meier method, evaluated via log-rank tests, indicated meaningful distinctions in IDFS between patients with low versus high TILs levels (using the 10% threshold) for the majority of models, although not universally. The models that reached significance delivered robust hazard ratios between 0.40 and 0.47 at this cutoff, apart from CellViT which showed a notably lower HR of 0.20.

Univariate Cox regression models applied to continuous TILs values—selected given the varied distributions noted in the analytical phase—yielded statistically significant associations for nine of the ten models **(Figure 6)**. Abousamra's model was the only one that did not (HR = 0.98, 95% CI = [0.97–1.0], p = 0.099). The nine significant models exhibited highly similar and overlapping hazard ratios: HoverNet HR = 0.91 (95% CI = [0.86–0.97], p = 0.004), CellViT HR = 0.93 (95% CI = [0.87–1.0], p = 0.048), KNN10 HR = 0.96 (95% CI = [0.93–0.99], p = 0.003), RT10 HR = 0.95 (95% CI = [0.93–0.98], p = 0.002), and NN10 HR = 0.96 (95% CI = [0.94–0.98], p < 0.001). Models trained on progressively larger datasets maintained comparable hazard ratios **(Figure 6)**, as seen with NN50 at HR = 0.96 (95% CI = [0.93–0.98], p = 0.002). Manual sTILs scoring by pathologists produced HR = 0.98 (95% CI = [0.97–1.00], p = 0.007) for reference.

Following adjustment for age group, tumor size group, histologic grade, and nodal status in multivariate models, outcomes remained largely consistent with the univariate findings **(Figure 7)**. CellViT and Abousamra's models were borderline non-significant after adjustment, recording HR = 0.94 (95% CI = [0.88–1.01], p = 0.097) and HR = 0.98 (95% CI = [0.97–1.00], p = 0.122), respectively. Within the chemotherapy-treated subset, adjusted hazard

ratios stayed similar across models for both continuous and categorical TILs variables. Significance was reached in every model except Abousamra's.

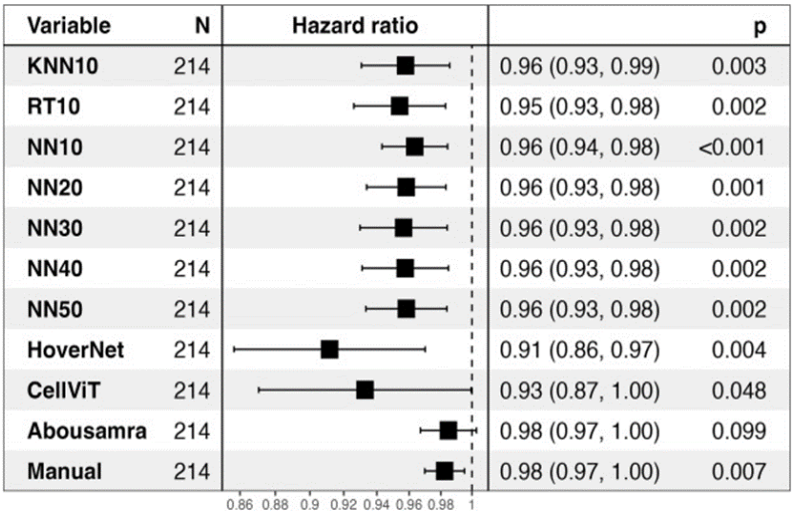


Figure 6. Forest plot presenting results from univariate Cox regression using continuous TILs scores from each method as predictors, with invasive disease-free survival (IDFS) as the outcome in the SCAN-B validation cohort.

Black squares mark the hazard ratio estimates, while horizontal error bars represent the 95% confidence intervals (CI). The CI values are additionally displayed in parentheses beside each hazard ratio. The scale along the bottom axis shows the range of hazard ratio values, and the vertical dashed line indicates the null value of HR = 1.

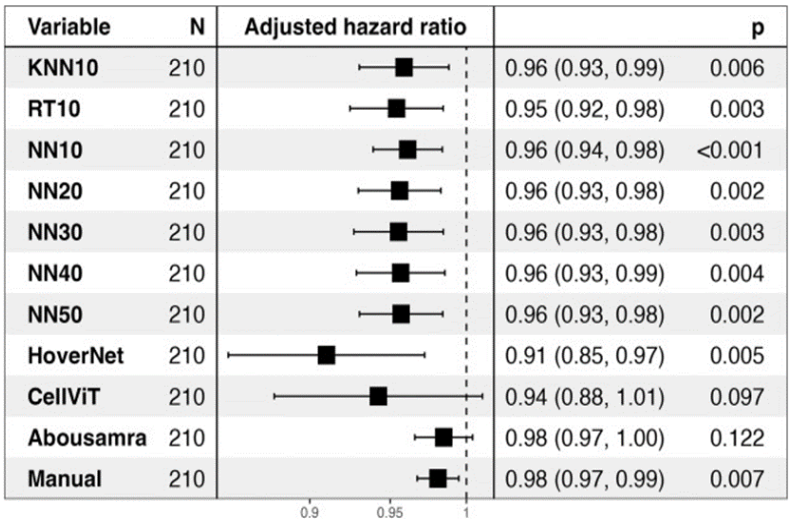


Figure 7. Forest plot from multivariate Cox regression of continuous TILs scores generated by each method (adjusted for age category, tumor size category, histologic grade, and nodal status), with invasive disease-free survival (IDFS) as the endpoint in the SCAN-B validation cohort.

Hazard ratios for the adjustment variables are omitted from the plot to save space. Black squares indicate the hazard ratio point estimates, horizontal error bars show the 95% confidence intervals (CI), and CI values are also listed in parentheses next to each HR. The numeric scale at the bottom represents the hazard ratio range, with the vertical dashed line marking the reference point of HR = 1.

Numerous AI systems designed for histopathology are currently available on the market, yet the broader integration of computational pathology into routine practice has progressed slowly. Although high-quality digital slides have existed for years and gained FDA approval in 2017, pathology departments have shown limited uptake. Manufacturers typically submit validation data to regulatory agencies to secure approvals such as FDA clearance or CE marking. However, conducting extensive trials that accurately reflect everyday clinical challenges remains

demanding. In recent years, various machine-learning strategies have been introduced to quantify anti-tumor immune activity, producing several TIL-related biomarkers with possible clinical relevance. Many of these AI-derived TIL markers, nevertheless, still lack the comprehensive validation required for routine use. There is also a shortage of comparative research examining multiple AI models with respect to both their analytical accuracy and prognostic strength.

This investigation evaluated the analytical and prognostic performance of ten AI TIL assessment tools against invasive disease-free survival in TNBC samples from the prospective Swedish SCAN-B study. Seven models (KNN, RT, and NN families) were developed using the QuPath platform for automatic TILs quantification. In addition, three established deep-learning CNN models—HoverNet [29], CellViT [30], and Abousamra *et al*. [31]—were incorporated as independent, high-performing reference predictors. Analysis of analytical performance revealed moderate to good agreement between different AI approaches, even when models were trained on substantial numbers of samples. Classifiers sharing similar architectures (NN10–NN50) exhibited strong internal consistency regardless of training size.

Clear differences in analytical validity appeared when comparing machine-generated TILs scores to manual pathologist sTILs evaluations in both internal and external validation sets. Within the internal Yale validation group, all AI models achieved good correlations with expert scores. Performance declined in the external SCAN-B set, resulting in only moderate correlations. Notably, expanding the training dataset size did not enhance agreement with manual scores. These findings point to meaningful discrepancies arising from variations in AI methodologies, training approaches, and especially differences in slide digitization platforms. The results underscore the necessity of rigorous testing across heterogeneous datasets and scanning systems to confirm the reliability and broad applicability of automated TILs tools.

A standard clinical 10% threshold [34–38] was applied to categorize TILs scores into low and high groups for all models. However, score distributions from most AI methods diverged substantially from those of pathologist sTILs. These differences stem from fundamental methodological distinctions. Manual sTILs assessment relies on semi-quantitative visual estimation of stromal area occupied by TILs, often using reference images, which can result in overestimation. In contrast, easTILs employs a precise mathematical calculation grounded in reproducible cell segmentation. Abousamra's patch-based method, which classifies entire image tiles rather than individual cells, aligns more closely with the visual approach of manual scoring and thus produces more similar distributions. Cell segmentation algorithms also play a role: watershed methods in QuPath tend to generate more false-positive detections by mistaking non-cellular structures for cells [29], while HoverNet and CellViT provide more accurate delineation, leading to smaller estimated cell areas and consequently lower TILs percentages via the easTILs formula [29]. Given these variations, continuous TILs scores are preferable over cutoff-based categories when comparing the prognostic performance of different AI models in survival analyses.

Despite observable differences in analytical metrics, the majority of AI models demonstrated strong prognostic value for digital TILs using IDFS as the endpoint, with the exception of CellViT [30] and Abousamra *et al*. [31]. Significant models produced consistent and overlapping hazard ratios across the full validation cohort as well as within the chemotherapy-treated subgroup. Notably, even models developed with relatively small training sets showed reliable prognostic capability. This outcome likely reflects the intrinsic strength of host anti-tumor immunity (as captured by TILs) as a biomarker, allowing adequately performing predictions even from less extensively trained systems. That said, reliance on very limited training data raises concerns about overfitting, which can restrict applicability in wider clinical settings. Therefore, developing such models on large, diverse datasets is essential to secure both robust prognostic performance and solid analytical generalizability.

One limitation of the current work is the absence of a regression-based model trained to directly predict TILs percentages, which would have broadened the range of methodologies examined. However, regression approaches might underperform compared with patch- or cell-classification strategies because they would inherit the known constraints and subjectivity of manual sTILs ground truth.

Successful clinical deployment of AI tools also demands explainable decision processes, especially when errors occur. Most models in this study rely on cell segmentation, offering superior interpretability over patch-based or regression methods. Pathologists and clinicians can visually inspect identified cells for potential misclassifications or missed objects, thereby increasing transparency and confidence in the system's outputs for real-world use.

Although the study did not assess inter-pathologist agreement on sTILs scoring due to the lack of overlapping evaluation subsets, this design choice actually strengthens the work by better simulating standard clinical conditions and incorporating natural scoring variability, as noted in the introduction.

The analysis of prospectively gathered data in a retrospective manner represents another constraint, as it does not directly demonstrate clinical utility. Level 1 evidence supporting the utility of manual TILs scoring has recently become available [39], but its impact on the acceptance or reimbursement of AI-based TILs assessment is yet to be determined. This study is among the earliest to perform a head-to-head comparison of analytical and prognostic validity for multiple independent AI TILs models in an external validation setting. Additional testing on the public TCGA dataset was not feasible, as it had already been used to pre-train some of the deep-learning models included here.

For clinical translation, these AI systems must undergo validation in representative, broad patient populations that reflect the real diversity of breast cancer cases. Because many existing studies use narrowly defined cohorts and some proprietary tools are not openly shared, creating standardized benchmarking frameworks would be valuable. Such frameworks could include a comprehensive, diverse reference dataset and neutral evaluation infrastructure that allows testing of both open-source and closed models without revealing proprietary details. Independent calculation of performance metrics would enhance trust, improve model transparency, and help clarify the strengths and weaknesses of each approach.

Conclusion

In conclusion, this work reveals notable variability in both analytical and prognostic validity (relative to invasive disease-free survival) among different AI TILs scoring systems when tested in an independent prospective cohort. There is an urgent requirement for a large, publicly accessible, multi-center benchmark dataset drawn from multiple clinical trials to thoroughly validate the prognostic usefulness of AI TILs tools and to promote fair comparisons across methodologies prior to clinical rollout. The ongoing multi-institutional CATALINA challenge study [40] may provide a suitable platform for this purpose.

Acknowledgments: None

Conflict of Interest: None

Financial Support: None

Ethics Statement: None

Research in context

Evidence before this study

Recent years have highlighted the growing relevance of tumor-infiltrating lymphocytes (TILs) for predicting outcomes and guiding therapy in early and metastatic triple-negative breast cancer (TNBC). Still, TILs scoring is a semi-quantitative process open to observer disagreement. Several machine-learning techniques have emerged to quantify anti-tumor immunity and reduce scoring inconsistency (PubMed search: TILs + artificial intelligence, 2024026). Although many AI tools focused on histopathology are commercially offered, integration of computational pathology into daily practice has been slow. Regulatory approvals rely on manufacturer-provided validation, but few independent studies replicate the complexities of actual clinical use. No previous work has systematically contrasted multiple AI models for both analytical agreement and prognostic accuracy.

Added value of this study

We document clear differences in analytical validity across ten AI TILs scoring systems when tested on internal and external patient groups. For prognostic validity, eight out of ten models reached statistical significance in a separate prospective cohort against invasive disease-free survival, producing comparable hazard ratios. Models trained with smaller sample sizes still achieved strong outcome prediction similar to those using larger datasets.

Implications of all the available evidence

The core robustness of the anti-tumor immune reaction (as reflected by TILs levels) appears sufficient for good performance even in models with modest training data. This same characteristic, however, raises the possibility of overfitting to particular training environments, which could limit broader usefulness. The results underscore the importance of establishing methods to ensure consistency and comparability across AI approaches before they enter clinical decision-making.

References

1. Burstein HJ, Curigliano G, Thürlimann B, Weber WP, Poortmans P, Regan MM, et al. Customizing local and systemic therapies for women with early breast cancer: the St. Gallen international consensus guidelines for treatment of early breast cancer 2021. Ann Oncol. 2021;32(10):1216-35.
2. Curigliano G, Burstein HJ, Gnant M, Loibl S, Cameron D, Regan MM, et al. Understanding breast cancer complexity to improve patient outcomes: the St Gallen international consensus conference for the primary therapy of individuals with early breast cancer 2023. Ann Oncol. 2023;34:970-86.
3. Denkert C, Loibl S, Noske A, Roller M, Müller BM, Komor M, et al. Tumor-associated lymphocytes as an independent predictor of response to neoadjuvant chemotherapy in breast cancer. J Clin Oncol. 2010;28(1):105-13.
4. Denkert C, von Minckwitz G, Darb-Esfahani S, Lederer B, Heppner BI, Weber KE, et al. Tumour-infiltrating lymphocytes and prognosis in different subtypes of breast cancer: a pooled analysis of 3771 patients treated with neoadjuvant therapy. Lancet Oncol. 2018;19(1):40-50.
5. Loi S, Drubay D, Adams S, Pruneri G, Francis PA, Lacroix-Triki M, et al. Tumor-infiltrating lymphocytes and prognosis: a pooled individual patient analysis of early-stage triple-negative breast cancers. J Clin Oncol. 2019;37:559-69.
6. Loi S, Michiels S, Adams S, Loibl S, Budczies J, Denkert C, et al. The journey of tumor-infiltrating lymphocytes as a biomarker in breast cancer: clinical utility in an era of checkpoint inhibition. Ann Oncol. 2021;32:1236-44.
7. Loi S, Sirtaine N, Piette F, Salgado R, Viale G, Van Eenoo F, et al. Prognostic and predictive value of tumor-infiltrating lymphocytes in a phase III randomized adjuvant breast cancer trial in node-positive breast cancer comparing the addition of docetaxel to doxorubicin with doxorubicin-based chemotherapy: BIG 02-98. J Clin Oncol. 2013;31(7):860-7.
8. Loi S, Drubay D, Adams S, Pruneri G, Francis PA, Lacroix-Triki M, et al. Abstract S1-03: pooled individual patient data analysis of stromal tumor infiltrating lymphocytes in primary triple negative breast cancer treated with anthracycline-based chemotherapy. Cancer Res. 2016;76(4_Supplement):S1-03.
9. Loi S, Salgado R, Adams S, Pruneri G, Francis PA, Lacroix-Triki M, et al. Tumor infiltrating lymphocyte stratification of prognostic staging of early-stage triple negative breast cancer. NPJ Breast Cancer. 2022;8(1):3.
10. Denkert C, Wienert S, Poterie A, Loibl S, Budczies J, Badve S, et al. Standardized evaluation of tumor-infiltrating lymphocytes in breast cancer: results of the ring studies of the international immuno-oncology biomarker working group. Mod Pathol. 2016;29(10):1155-64.
11. Salgado R, Denkert C, Demaria S, Sirtaine N, Klauschen F, Pruneri G, et al. The evaluation of tumor-infiltrating lymphocytes (TILS) in breast cancer: recommendations by an international TILS working group 2014. Ann Oncol. 2015;26:259-71.
12. Hendry S, Salgado R, Gevaert T, Russell PA, John T, Thapa B, et al. Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the international immunooncology biomarkers working group: part 1: assessing the host immune response, TILs in invasive breast carcinoma and ductal carcinoma in situ, metastatic tumor deposits and areas for further research. Adv Anat Pathol. 2017;24:235-51.
13. Choi S, Cho SI, Jung W, Jin I, Lee A, Kim H, et al. Deep learning model improves tumor-infiltrating lymphocyte evaluation and therapeutic response prediction in breast cancer. NPJ Breast Cancer. 2023;9(1):71.
14. Bai Y, Cole K, Martinez-Morilla S, Elghazawy M, Hathaway M, Brcic I, et al. An open source, automated tumor infiltrating lymphocyte algorithm for prognosis in triple-negative breast cancer. Clin Cancer Res. 2021;27(20):5557-65.
15. Matikas A, Johansson H, Grybäck P, Bjöhle J, Acs B, Holmberg E, et al. Survival outcomes, digital TILs, and on-treatment PET/CT during neoadjuvant therapy for HER2-positive breast cancer: results from the randomized PREDIX HER2 trial. Clin Cancer Res. 2023;29(3):532-40.
16. Yuan Y. Modelling the spatial heterogeneity and molecular correlates of lymphocytic infiltration in triple-negative breast cancer. J R Soc Interface. 2015;12(103):20141153.

17. Klauschen F, Müller KR, Binder A, Bockmayr M, Hägele M, Seegerer P, et al. Scoring of tumor-infiltrating lymphocytes: from visual estimation to machine learning. Semin Cancer Biol. 2018;52:151-7.
18. Bera K, Schalper KA, Rimm DL, Velcheti V, Madabhushi A. Artificial intelligence in digital pathology — new tools for diagnosis and precision oncology. Nat Rev Clin Oncol. 2019;16(11):703-15.
19. Ibrahim A, Gamble P, Jaroensri R, Abdelsamea MM, Mermel CH, Chen PC, et al. Artificial intelligence in digital breast pathology: techniques and applications. Breast. 2020;49:267-73.
20. 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.
21. Amgad M, Stovgaard ES, Balslev E, Thagaard J, Chen W, Dudgeon S, et al. Report on computational assessment of tumor infiltrating lymphocytes from the international immuno-oncology biomarker working group. NPJ Breast Cancer. 2020;6:16.
22. Reisenbichler ES, Han G, Bellizzi A, Bossuyt V, Brock J, Cole K, et al. Prospective multi-institutional evaluation of pathologist assessment of PD-L1 assays for patient selection in triple negative breast cancer. Mod Pathol. 2020;33(9):1746-52.
23. Rydén L, Loman N, Larsson C, Hegardt C, Vallon-Christersson J, Malmberg M, et al. Minimizing inequality in access to precision medicine in breast cancer by real-time population-based molecular analysis in the SCAN-B initiative. Br J Surg. 2018;105(2):e158-68.
24. Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, et al. The Sweden Cancerome Analysis Network - breast (SCAN-B) initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Med. 2015;7(1):20.
25. Staaf J, Glodzik D, Bosch A, Vallon-Christersson J, Reuterswärd C, Häkkinen J, et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. Nat Med. 2019;25(10):1526-33.
26. Dent R, Trudeau M, Pritchard KI, Hanna WM, Kahn HK, Sawka CA, et al. Triple-negative breast cancer: clinical features and patterns of recurrence. Clin Cancer Res. 2007;13(15):4429-34.
27. Acs B, Fredriksson I, Rönnlund C, Hober S, Ehinger A, Rydén L, et al. Variability in breast cancer biomarker assessment and the effect on oncological treatment decisions: a nationwide 5-year population-based study. Cancers. 2021;13(5):1-15.
28. Bankhead P, Loughrey MB, Fernández JA, Dombrowski Y, McArt DG, Dunne PD, et al. QuPath: open source software for digital pathology image analysis. Sci Rep. 2017;7(1):16878.
29. Graham S, Vu QD, Raza SEA, Azam A, Tsang YW, Kwak JT, et al. Hover-Net: simultaneous segmentation and classification of nuclei in multi-tissue histology images. Med Image Anal. 2019;58:101563.
30. Hörst F, Rempe M, Heine L, Seibold C, Keyl J, Baldini G, et al. CellViT: vision transformers for precise cell segmentation and classification. arXiv [Preprint]. 2023 [cited 2024 Apr 4]. Available from: http://arxiv.org/abs/2306.15350
31. Abousamra S, Gupta R, Hou L, Batiste R, Zhao T, Shankar A, et al. Deep learning-based mapping of tumor infiltrating lymphocytes in whole slide images of 23 types of cancer. Front Oncol. 2022;11:806603.
32. Gamper J, Alemi Koohbanani N, Benet K, Khuram A, Rajpoot N. PanNuke: an open pan-cancer histology dataset for nuclei instance segmentation and classification. In: Digital pathology: 15th European congress, ECDP 2019, Warwick, UK, April 10–13, 2019, proceedings 15. Springer; 2019. p. 11-9.
33. Hudis CA, Barlow WE, Costantino JP, Gray RJ, Pritchard KI, Chapman JA, et al. Proposal for standardized definitions for efficacy end points in adjuvant breast cancer trials: the STEEP system. J Clin Oncol. 2007;25(15):2127-32.
34. Asano Y, Kashiwagi S, Goto W, Takada K, Takahashi K, Hatano T, et al. Prediction of treatment response to neoadjuvant chemotherapy in breast cancer by subtype using tumor-infiltrating lymphocytes. Anticancer Res. 2018;38(4):2311-21.
35. Goto W, Kashiwagi S, Asano Y, Takada K, Takahashi K, Hatano T, et al. Predictive value of improvement in the immune tumour microenvironment in patients with breast cancer treated with neoadjuvant chemotherapy. ESMO Open. 2018;3(6):e000305.
36. Jang N, Kwon HJ, Park MH, Kang SH, Bae YK, Lee SJ, et al. Prognostic value of tumor-infiltrating lymphocyte density assessed using a standardized method based on molecular subtypes and adjuvant chemotherapy in invasive breast cancer. Ann Surg Oncol. 2018;25(4):937-46.

37. Ruan M, Tian T, Rao J, Xu X, Yu B, Yang W, et al. Predictive value of tumor-infiltrating lymphocytes to pathological complete response in neoadjuvant treated triple-negative breast cancers. Diagn Pathol. 2018;13(1):66.
38. Park HS, Heo I, Kim JY, Kim S, Nam S, Park S, et al. No effect of tumor-infiltrating lymphocytes (TILs) on prognosis in patients with early triple-negative breast cancer: validation of recommendations by the international TILs working group 2014. J Surg Oncol. 2016;114(1):17-21.
39. Leon-Ferre RA, Jonas SF, Salgado R, Loi S, De Schepper M, Punchai S, et al. Tumor-infiltrating lymphocytes in triple-negative breast cancer. JAMA. 2024;331(13):1135.
40. CATALINA-Challenge (CollAborative Til vALidatIoN chAllenge) [Internet]. [cited 2024 Apr 4]. Available from: https://www.tilsinbreastcancer.org/tils-grand-challenge/