A Machine Learning Model Optimized for Local Data Stratifies Patients for the Adoptive Cell Therapy with Tumor Infiltrating Lymphocytes in Bladder Tumors

Kayode D. Olumoyin[1], Ahmet Murat Aydin[2], Sarah Bazargan[3], Brittany Bunch[3], Ibrahim Chamseddine[1], Aleksandra Karolak[4], Matthew Beatty[3], Shari Pilon-Thomas[3], Michael A. Poch[2], Katarzyna A. Rejniak[1,5,*]

[1]Integrated Mathematical Oncology, H. Lee Moffitt Cancer Center & Research Institute
[2]Department of Genitourinary Oncology, H. Lee Moffitt Cancer Center & Research Institute
[3]Department of Immunology, H. Lee Moffitt Cancer Center & Research Institute
[4]Department of Machine Learning, H. Lee Moffitt Cancer Center & Research Institute
[5]University of South Florida, Morsani College of Medicine, Department of Oncologic Sciences

ABSTRACT

Adoptive cell therapy with tumor-infiltrating lymphocytes (ACT-TILs) involves autologous TILs that are expanded ex vivo and then reinfused into the patient as a personalized form of therapy. For bladder cancer patients, there is a unique opportunity to deliver TILs locally, through intravesical administration, without systemic cytotoxic chemotherapy to induce lymphodepletion. Improved selection of patients for ACT-TILs by predicting TIL expansion may prevent excessive treatment-related costs and delays in treatment of patients whose bladder cancer specimens would not yield successful TIL growth. In this study, we developed a machine-learning (ML) model optimized for local data of a medium size to determine a minimal robust feature combination (demographic, clinical, and biological tumor specimen-based) that is predictive of whether TILs can be expanded from a resected bladder cancer. The study uses a retrospectively identified set of data from bladder cancer patients at Moffitt Cancer Center that were assigned to internal training/testing or blinded validation cohorts. Using random forest method (RF) to identify a combination of robust predictive features, support vector machine (SVM) model to determine the optimal classification hyperparameters, and Matthews correlation coefficient (MCC) method to adjust the decision-boundary threshold for imbalanced data, our model yielded AUC=0.771 for the internal testing cohort and AUC=0.828 for the blinded validation cohort. Thus, our ML model optimized for local data of medium size has favorable performance metrics for predicting TIL expansion from a given tumor. This computational predictor can serve as a clinical supportive tool to determine which patients are candidates for ACT-TILs.

Introduction

Bladder cancer is the fourth most common cancer among men and a leading cause of cancer death among men and women. In the United States in 2024, approximately 83,190 new cases of bladder cancer were estimated with 16,840 bladder cancer deaths [1]. Despite current therapies, 50% of patients with intermediate and high-risk localized non-muscle invasive bladder cancer fail bladder-sparing treatment. Additionally, recurrent or locally advanced tumors failing bladder-sparing treatment have an even worse prognosis, often requiring radical cystectomy, a radical and lengthy operation, which is associated with several co-morbidity and changes in quality-of-life afterwards [2,3]. Further research using novel approaches is needed to treat this disease at every stage.

One major advance for treating solid tumors is the success of adoptive cell therapy (ACT) during which autologous tumor-infiltrating lymphocytes (TILs) are expanded and activated ex vivo and then reinfused into the cancer patient [4]. Indeed, ACT with TIL has emerged as one of the most powerful therapies for unresectable metastatic melanoma and cervical cancer [5-7]. Furthermore, it has been recently shown that ACT-TIL provided longer progression-free survival than the immune checkpoint inhibitors among patients with advanced melanoma [8]. Because bladder tumors have a high mutational burden corresponding to an increased number of neoantigens [9,10], these cancer cells could be recognized by activated T cells at the tumor site, making the bladder cancer a potential candidate for ACT-TIL treatment. Prior studies by us and others showed that TILs expanded from bladder cancer recognize autologous tumor [11], and that expanding TILs ex vivo from the resected bladder tumors is feasible [12,13]. Moreover, translating ACT-TIL in bladder cancer provides a unique opportunity to deliver TIL intravesically by administering T cells through a catheter into the bladder directly to tumors. A phase I clinical trial testing the safety and feasibility of intravesical administration of ACT-TIL among patients with high-grade non-muscle invasive bladder cancer (NMIBC) is currently open at Moffitt Cancer Center (NCT05768347) [14].

One of the main steps in the ACT-TIL approach is the ability to expand tumor-reactive TIL to large amounts for reinfusion into patients' bladders. Our previous work [13] showed that about 70% of collected primary bladder tumors were capable of TIL expansion. However, this process takes several weeks. Therefore, it would be beneficial to prognosticate early whether a patient will benefit from TIL therapy to allow for better patient stratification. One approach is using quantitative methods, such as machine-learning (ML) algorithms, to predict whether TILs can be expanded from resected tumors based on patient data. Machine-learning methods have been previously applied to several aspects of bladder cancer, such as evaluation of cancer stage in computed tomography (CT) urography [15]; predictions of early recurrence of non-muscle invasive bladder cancer (NMIBC) based on histology images [16]; and stratification of muscle-invasive bladder cancer (MIBC) patients into two survival groups (*good* and *poor*) after chemotherapy [17]. Here, we present the **M**achine **L**earning **P**redictor of the **E**xpansion of **T**umor **I**nfiltrating **L**ymphocytes (ML-PETIL). ML-PETIL is a tool that can first learn from patient and tumor data already collected in the clinic (local data) which data features are important for predicting TIL expansion, without the need to predefine which data categories to consider. Then, this tool predicts a possible TIL expansion for individual patients (personalized predictions) allowing to determine whether ACT-TIL therapy could potentially treat an individual bladder cancer patient.

Results

The goal of our studies was twofold. First, we aimed to identify a combination of data features among those previously collected in the clinic (and grouped as demographic, clinical, and specimen-based data) that were predictive of whether TIL can be expanded from that tumor. In this way, our model is optimized for local data, that is in contrast to other methods that use the predefined categories for data classification. Next, we used those data features to create and validate an ML predictor, ML-PETIL, to stratify individual patients into *Yes-TIL* vs. *No-TIL* classes that define TIL expansion potential.

Study design

The prospective database of 102 adult patients was created between 2015 and 2022 by collecting clinicopathologic and specimen information of bladder cancer patients undergoing surgery at Moffitt Cancer Center (Table 1). Data collection protocols were approved by the Advarra Institutional Review Board (MCC18142 and MCC20106). Informed consent was obtained from all patients prior to data collection. Among the collected datasets, 68 tumors demonstrated TIL

expansion (class *Yes-TIL*), and 34 did not (class *No-TIL*). The initial screening identified the maximal subset of 60 patients with complete information on 16 commonly collected features that included (i) patients' demographics: age at surgery, body-mass index (BMI), race, and smoker status; (ii) tumor clinical characteristics: clinical stage (cT), pathological stage (pT), pathological lymph node stage (pN), type of surgery, surgeon, prior neoadjuvant chemotherapy (NAC), type of radical cystectomy histology (Histology), type of cystoscopic biopsy histology (Bx Histology), and cT/pT status; and (iii) specimen measurements: tumor sample weight, number of fragments plated for TIL expansion, and primary tumor digest count. The remaining patients were missing data values in some of the considered features. Because individual features were not predictive of the TIL expansion status (Pearson correlation coefficients were between −0.22444 and 0.31394 for each of the 16 features, Supplementary Table 1), the combinations of features were considered, and the ML approach was applied. We tested both the complete case analysis approach and the case with imputed data using the Multiple Imputation with Denoising Autoencoders (MIDAS)[18,19] method. However, because the data features were both numerical and categorical, the data imputation did not improve overall predictability of our model (Supplementary Figure 1).

Table 1: Description of Study Population

Features	Number of missing data[a]	Total Cohort (N=102)
Age at Surgery, mean ± SD	0	69.0 ± 11.6
Race, n (%)	0	
American Indian / Alaskan Native		1 (1.0%)
Asian		1 (1.0%)
Black		2 (2.0%)
White		97 (95.1%)
Unknown		1 (1.0%)
Surgery, n (%)	0	
Radical Cystectomy		70 (68.6%)
TURBT		32 (31.4%)
Smoker, n (%)	0	
Current Smoker		14 (13.7%)
Ever Smoked		56 (54.9%)
Never Smoked		22 (21.6%)
Not Reported		10 (9.8%)
BMI, mean ± SD	0	29.3 ± 7.6
NAC, n (%)	3	
NAC		39 (39.4%)
No NAC		60 (60.6%)
cT, n (%)	9	
Ta		17 (18.3%)
CIS		2 (2.2%)
T1		17 (18.3%)
T2		39 (41.9%)
T3		14 (15.1%)
T4		4 (4.3%)
pT, n (%)	21	
T0		5 (6.2%)
Ta		7 (8.6%)
CIS		3 (3.7%)
T1		4 (4.9%)
T2		20 (24.7%)
T3		28 (34.6%)
T4		14 (17.3%)
pN, n (%)	32	
N0		48 (68.6%)
N1		9 (12.9%)
N2		5 (7.1%)

N3		6 (8.6%)
Nx		2 (2.9%)
Bx Histology, n (%)	2	
Benign		2 (2.0%)
Urothelial Carcinoma		95 (95.0%)
Squamous Cell Carcinoma		2 (2.0%)
Small Cell Carcinoma		1 (1.0%)
Histology, n (%)	3	
Benign		2 (2.0%)
Urothelial Carcinoma		95 (96.0%)
Squamous Cell Carcinoma		2 (2.0%)
Sample weight (g), mean ± SD	0	1.1 ± 2.2
Number of Fragments plated (tumor), mean ± SD	0	13.6 ± 8.8
Primary tumor digest count, mean ± SD	0	$1.47 \times 10^7 \pm 3.44 \times 10^7$

SD: standard deviation; BMI: body mass index; cT: clinical tumor stage; pT: pathological tumor stage; pN: pathological lymph node stage; Bx Histology: biopsy histology; NAC: Neoadjuvant chemotherapy. TNM staging system for primary tumor: T0, no evidence of primary tumor; Ta, noninvasive papillary carcinoma; CIS, carcinoma in situ; T1, T2, T3, T4, size and/or extension of the primary tumor. TNM staging system for lymph nodes: N0, no regional lymph node metastasis; N1, metastasis in one lymph node; N2, Metastasis in multiple lymph nodes; N3, metastasis in lymph nodes along the common iliac vessels; Nx, regional lymph nodes not assessed. [a]The number of patients missing data in at least one of the 16 clinical, demographical, and specimen information features.

Therefore, we used the complete dataset with 60 patients (47 from a *Yes-TIL* class and 13 from a *No-TIL*) for our study. This data was randomly assigned to the internal cohort (n = 40) for determining the robust predictive data features or to the blinded cohort (n = 20) for final validation step. For every patient, information was recorded on whether the resected tumor sample caused positive TIL growth. This information was used only for the internal cohort. For the final validation cohort, the TIL growth status was blinded until after the ML-PETIL predictor was developed and the predictions for this cohort were generated. The blinded validation cohort was created due to the lack of an external validation dataset from another institution, because the dataset we used here is the biggest collection of bladder cancer TIL specimens in one institution.

The 40 patients' datasets in the internal cohort were subjected to the random forest (RF) algorithm to determine features that were robust in their predictions. As a result, 7 features (out of 16) were identified as robust. Next, all initial datasets not selected as the blinded validation cohort was considered, and data containing all 7 robust features formed the training/testing cohort, which was used to build the ML-PETIL predictor and carry its initial testing. This cohort comprised a total of 82 patients with no missing data in the 7 features. Subsequently, this dataset was split 70−30 into a training dataset with 57 patients and a testing dataset with 25 patients. The training cohort was used to learn the optimal hyperparameters of the support vector machine (SVM) algorithm. Next, the Matthews correlation coefficient (MCC) method was used to adjust the optimal decision-boundary threshold to account for the imbalance in *Yes-TIL* and *No-TIL* data in the training cohort. This included a cross-validation step. These steps constituted the development of the ML-PETIL model that was subsequently used to stratify patients in the testing cohort. Finally, the stratification was performed with the additional blinded validation cohort. For each cohort, the performance metrics were computed, including accuracy, sensitivity, and specificity. The detailed flowchart for this study is shown in Figure 1.

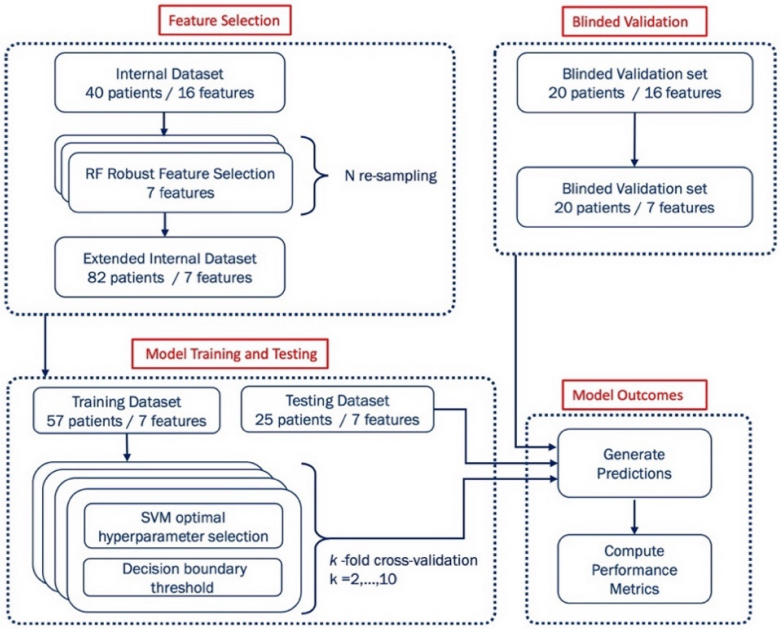

Figure 1. The ML-PETIL flowchart for this study. Abbreviation: ML-PETIL, Machine Learning Predictor of the Expansion of Tumor Infiltrating Lymphocytes.

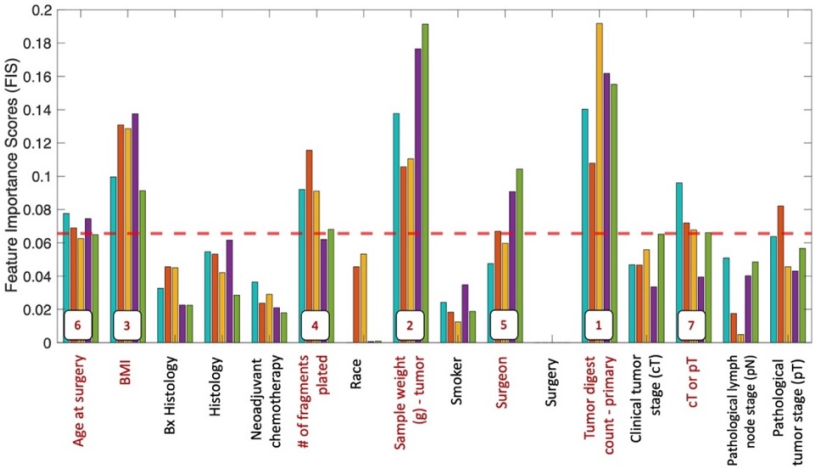

Figure 2. Selection of robust predictive features. The FIS values for 5 data subsamplings with the selection threshold shown as a dashed red line. The 7 robust predictive features identified by the RF algorithm are numbered in the order of their average FIS value. Abbreviation: FIS, feature importance score; RF, random forest.

Identifying a combination of robust predictive features

The 16 features commonly collected in the clinic were used as a base for identifying a smaller set of robust predictive features. The base set included 6 demographic, 7 clinical, and 3 biological tumor specimen-based features, all listed in the Materials and Methods section. Starting with an internal cohort of 40 patient datasets, the random forest (RF) [20,21] algorithm was used to resample 70% of the training cohort data 5 times and the feature importance score (FIS) [20,22,23] for each of the 16 features was calculated (Supplementary Figure 2). The selection criterium for robust

predictive features required that at least 3 out of 5 calculated FIS values were at least 5% above the probability that every feature is equally likely (the selection threshold is 0.065625 in this case). The RF algorithm determined 7 robust predictive features: (i) two demographic features (age at surgery, BMI), (ii) two clinical features (surgeon, cT/pT); and (iii) three biological tumor specimen-based features (number of fragments plated, sample weight tumor, primary tumor digest count) shown in Figure 2. The distributions of the robust predictive feature values in all 3 cohorts are shown in Supplementary Figure 3**.**

Local data-derived learning of the predictive SVM and MCC hyperparameters

For the 7 robust predictive features, a total of 82 patients' datasets were missing no data. This cohort was split 70-30 into a training dataset of 57 patients (36 *Yes-TIL* and 21 *No-TIL*) that was used in a learning phase of the classifier accompanied with the internal cross-validation (CV), and a testing dataset of 25 patients (16 *Yes-TIL* and 9 *No-TIL*) that was used for testing classifier predictivity. To capture interactions between predictive feature values that are potentially multidimensional and nonlinear, the support vector machine (SVM) model [24,25] was used to learn the optimal classification hyperparameters based on the training cohort. Four SVMs with different kernels were tested, of which the radial basis kernel function (RBF) performed better than polynomial, linear, and sigmoid kernels (Supplementary Figure 4).

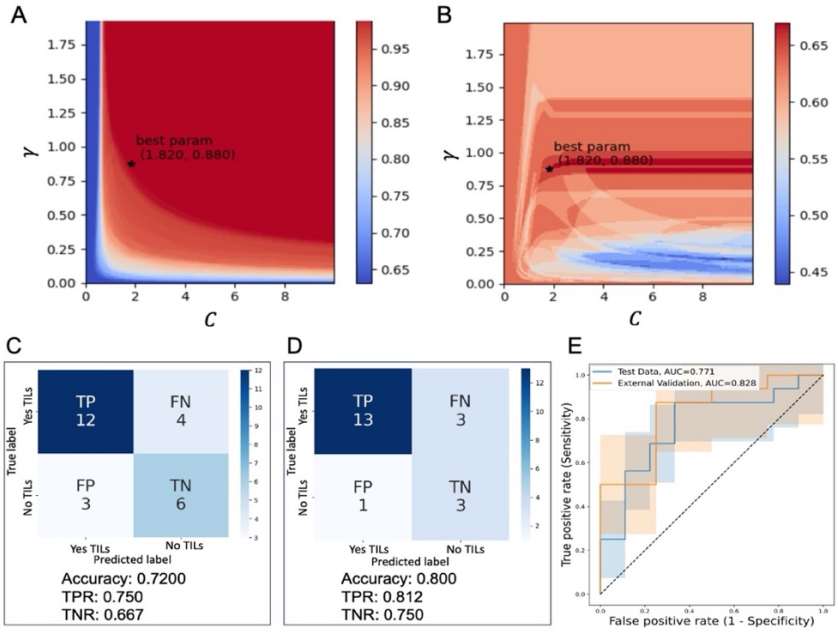

Figure 3. Identification of the predictive SVM hyperparameters and model performance analysis. The heatmaps for an optimal k-fold CV of k=8 show the average predicted accuracies for **A.** k-1 training partitions and **B.** 1 testing partition. The optimal SVM hyperparameters C and γ (*best param*) were identified by simultaneously (i) maximizing the accuracy value for training partitions, (ii) minimizing the difference between the accuracies in training and testing partitions, and (iii) minimizing the value of C. Colors indicate the accuracy value. **C.** The confusion matrix of ML-PETIL predictions for the testing dataset with the indicated performance metrics: the accuracy, the TPR, and the TNR. **D.** The confusion matrix of ML-PETIL predictions for the blinded validation set. **E.** ROC curves generated for ML-PETIL predictions for the testing (blue) and blinded validation (orange) cohorts with the indicated confidence intervals. The dashed line represents a random classifier. Abbreviations: ML-PETIL, Machine Learning Predictor of the Expansion of Tumor Infiltrating Lymphocytes; ROC, receiver-operating characteristic; TNR, true negative rate; TPR, true positive rate.

The optimal RBF-SVM hyperparameters (C, γ) were identified by first, performing k-fold CV on the training dataset (for k = 2, …,10) and then by computing the Matthews correlation coefficient (MCC) for each cross-validation to account for unbalanced data. Here, C specifies the width of the margins for avoiding data misclassification, while γ controls the nonlinearity of the decision boundary hyperplane. For each k, the best C and γ were determined by (i) minimizing the difference between the cross-validation training accuracy and testing accuracy; (ii) maximizing the cross-validation training accuracy; and (iii) minimizing the value of C. The MCC values ranged from 0.632 to 1.0, with the optimal value of 1.0 obtained for CV of k = 4, 6, 8, and 10. In these cases, the obtained hyperparameters values were: C = 2.390 for k = 4; C = 2.520 for k = 6; C = 1.820 for k = 8, and C = 3.030 for k = 10, which yielded the minimal C of 1.820 for k = 8 and γ = 0.88 (these are denoted as *best param* in Figure 3A,B). For this case, the decision boundary threshold that yielded MCC = 1.0 was 0.606. Analysis of the other hyperparameter cases is shown in Supplementary Figure 5, and the optimal performance of the testing and blinded validation set of the chosen hyperparameters is presented in Supplementary Figure 6.

Evaluating the predictive performance of ML-PETIL

The developed optimal SVM-MCC model was first verified using the testing dataset and was subsequently validated using the blinded validation dataset. The testing cohort comprised 25 patients, and the ML-PETIL predictions yielded the accuracy of 0.720, the true negative rate (specificity) of 0.667, and the true positive rate (sensitivity) of 0.750 (Figure 3C). The validation cohort comprised 20 patients (with the *Yes-TIL* and *No-TIL* classifications blinded) and generated an accuracy of 0.800, a specificity of 0.750, and a sensitivity of 0.812 (Figure 3D). The area under the curve (AUC) for the testing cohort was 0.771, and the AUC for the blinded validation cohort was 0. 828 (Figure 3E).

Clinical application of ML-PETIL

Because ex vivo expansion of TIL takes about 4 weeks after tumor resection, early assessment of the outcome of TIL expansion could benefit the patient. The proposed ML model is optimized for local data and includes features derived from the already collected information. With this model, we report the probability of tumor classification into *Yes-TIL* and *No-TIL* categories. The ML-PETIL tool can help a clinician to consider ACT-TIL if a patient is predicted to have TIL expansion. In contrast, if a patient is predicted to have a negative TIL expansion, other therapies could be considered. This improved selection of patients for ACT-TILs may prevent excessive treatment-related costs and delays in treatment of patients whose bladder cancer specimens would not yield successful TIL growth. The ML-PETIL predictor is a practical model of patients' stratification, optimization of treatment strategies, and guidance of clinical decision-making.

Discussion

Despite recent advancements in treatment of non-muscle invasive bladder cancer [26], especially once patients fail BCG (Bacillus Calmette-Guerin) a standard intravesical immunotherapy, long-term results are modest, with low durable treatment response rates and cystectomy-free rates far from ideal [27,28]. These modest results show a significant unmet need for further development of therapies for bladder cancer patients, preferably treatments that can be delivered locally with minimal associated adverse events and that can provide durable responses. Because bladder tumors have a high mutational burden, they are good candidates for adoptive cell therapy with autologous tumor-infiltrating lymphocytes (ACT-TIL), in which TILs are expanded and activated *ex vivo,* and then reinfused into the cancer patient as a personalized form of therapy. For bladder cancer patients, there is also a unique opportunity to deliver TILs locally, through intravesical administration, without systemic cytotoxic chemotherapy to induce lymphodepletion. Although ACT with TIL holds a great promise, we have previously shown that TIL expansion from the

resected bladder tumors failed in about 30% of the patients [13]. Thus, early identification of patients who could benefit from ACT-TIL is essential for optimizing treatment strategies and in preventing delays in next appropriate treatment.

We conducted a retrospective analysis of multimodal data of 102 bladder cancer patient specimens and developed the ML-PETIL model for predicting whether a given tumor will be suitable for the successful expansion of TILs. One of the important features of this method is the capacity to first identify which data combinations are important for making predictions from a larger dataset in hand that has been already collected in clinic. Thus, our method does not require a predefined set of data features, but it learns which data features are robust in predicting the outcome from the available data. As a result, this method is optimized for the local data. Once the combination of robust predictive features was identified, the ML-PETIL predictor showed favorable performance metrics on both the testing cohort (AUC = 0.771) and the blinded validation cohort (AUC = 0.828).

Another important feature of ML-PETIL model is the capacity to make successful predictions using cohorts of a medium size (57 datasets for a training cohort, 25 for a testing cohort, and 20 for a blinded validation cohort). This is in contrast to other ML methods, and deep learning methods in particular, that require large datasets for training. In our studies, some patients were excluded because of missing data in the considered features. While several imputation methods were tested, the overall predictability of the model with imputed data was not improved, because the data features were both numerical and categorical. Another limitation of this study is the lack of an external validation dataset from another institution. We created the validation cohort from the initial data where the *Yes-TIL*/*No-TIL* classifications were blinded until after the ML-PETIL model generated predictions for this validation cohort. We expect, that when the larger datasets become available, ML-PETIL may reach even better predictability. To our knowledge, the dataset we used is the biggest collection of bladder cancer TIL specimens in one institution, and we are continuing to accrue bladder tumors from patients at Moffitt Cancer Center to include in future studies [12,13].

In conclusion, a novel ML model was developed based entirely on local data already collected in clinic without the need of acquiring additional data features or algorithm-based imputation. This model was able to predict TIL expansion with favorable performance. ML-PETIL computational predictor can be used as a clinical-supportive tool to stratify patient eligibility for immune-based therapies, such as the ACT-TIL. Such improved selection may prevent excessive treatment-related costs and delays in treatment of patients whose bladder cancer specimens would not yield successful TIL growth.

Methods

Clinical and experimental data

The prospective database of 102 adult patients was created between 2015 and 2022 by collecting clinicopathologic and specimen information of bladder cancer patients undergoing surgery at Moffitt Cancer Center. Data collection protocols and all experimental protocols were approved by the Advarra Institutional Review Board (MCC18142 and MCC20106) and performed in accordance to IRB guidelines. Informed consent was obtained from all patients prior to data collection. Bladder tumors were included if they were larger than 1 cm and if there was tissue specimen available after pathological diagnosis. Tumors specimens were collected from transurethral resection of bladder tumor (TURBT) or radical cystectomy. Tumor infiltrating lymphocytes were expanded from resected surgical tumors as previously described [13]. Tumors were minced into fragments, placed in individual wells of a 24-well plate, and propagated in media containing 6000 IU/mL IL-2 for up to four weeks. As cell confluence was reached, each well was split into additional wells. Expansion of TIL was considered successful if at least 2 wells were

confluent at the end of four weeks. These cases were categorized as *Yes-TIL*, and all other cases as *No-TIL*. This database was used for retrospective analysis and ML-based predictions.

The initial screening identified the maximal subset of 60 patients with complete information on 16 commonly collected features that included information about patients' (i) demographics: patient's age at surgery, body-mass index (BMI), race, and smoker status; (ii) clinical characteristics: clinical tumor stage (cT), pathological tumor stage (pT), pathological lymph node stage, type of surgery, surgeon, prior neoadjuvant chemotherapy, type of radical cystectomy histology, type of cystoscopic biopsy histology, and cT/pT status; and (iii) specimen information: tumor sample weight, number of fragments plated for TIL expansion, and primary tumor digest count. The patients excluded from this initial stage of our model were missing data in some considered features. We used the complete case analysis approach, since the data features were both numerical and categorical, and the imputation methods that were tested did not improve overall predictability of our model. This data was randomly assigned to the internal cohort (n = 40) for determining the robust predictive data features or to the blinded cohort (n = 20) for final validation step. For every patient, information was recorded on whether the resected tumor sample caused positive TIL growth. This information was used only for the internal cohort. For the final validation cohort, the TIL growth status was blinded until after the ML-PETIL predictor was developed and the predictions for this cohort were generated. The blinded validation cohort was created due to the lack of an external validation dataset from another institution, since the dataset we used here is the biggest collection of bladder cancer TIL specimens in one institution.

Random Forest method for identifying robust predictive features

The random forest (RF) algorithm [20,21] is an ensemble-learning method that combines outputs of multiple decision trees to select the majority-based decision. Random forest was applied to the internal cohort to identify which locally collected data features form a robust predictive feature subset. This identification was done by resampling 70% of the training cohort data 5 times, and generating a feature importance score (FIS) [20,22,23] for each feature. A detailed algorithm is presented in Supplementary Figure 7. The selection criterium for a robust predictive feature required that its FIS value be at least 5% above the uniform threshold in at least 3 of the 5 resamples of the internal dataset. The uniform threshold was defined as the probability that all features have equal likelihood of being predictive.

Support Vector Machine model for learning the optimal classification rules

All data from the initial dataset that contain the identified robust predictive features and were not assigned to the blinded validation cohort were used for learning the optimal classification rules. This data was split 70-30 into the training and testing cohorts. To capture interactions between predictive feature values that are potentially multidimensional and nonlinear, the support vector machine (SVM) model [24,25] was used to learn the optimal classification hyperparameters based on the training cohort. Four SVMs with different kernels were tested: linear, polynomial, sigmoid, and radial basis function. In a binary classification task under consideration, the dataset $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_M]^{\mathsf{T}}$ consists of M data points $\boldsymbol{X}_i = \left(\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^P\right)$ with P features and the corresponding binary target class $y_i \in \{-1, 1\}$ for each i. The linear SVM generates a predicted binary class $\widehat{y_i} \in \{-1, 1\}$ for each data point $\boldsymbol{X}_i$ that is assigned according to the following rule: $\widehat{y_i} = -1 \; for \;\; \boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}_i + b < 0$, and $\widehat{y_i} = 1 \; for \;\; \boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}_i + b > 0$. The objective function that is minimized during the SVM training is defined as: $\min_{\boldsymbol{w},b} \left\{\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{M} H(y_i, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}_i + b\,)\right\}$, where $H(y_i, \boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}_i + b\,) = max(0, 1 - y_i \times \boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}_i + b)$ is the convex hinge loss [29]. The parameter $\boldsymbol{w}$ is the learnable weight vector, b is the bias value, C specifies the width of the margins for avoiding data misclassification. The linear SVM is extended to a nonlinear SVM by choosing an appropriate kernel function $K(.,.)$ that transforms each $\boldsymbol{X}_i$ into a higher dimensional space. We used the

following kernel functions: linear $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = (\boldsymbol{X}_i{}^{\mathsf{T}}\boldsymbol{X}_j)^2$, polynomial $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = (\gamma \boldsymbol{X}_i{}^{\mathsf{T}}\boldsymbol{X}_j + r)^N$, sigmoid $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = \tan(\gamma \boldsymbol{X}_i{}^{\mathsf{T}}\boldsymbol{X}_j + r)$, and radial basis function $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = exp\left(-\gamma \|\boldsymbol{X}_i - \boldsymbol{X}_j\|^2\right)$, where γ controls the nonlinearity of the decision boundary hyperplane. The SVM with a radial basis kernel function (RBF-SVM) performed better than polynomial, sigmoid and linear kernels. The optimal RBF-SVM hyperparameters (C, γ) were identified by performing k-fold cross-validation (CV) on the training dataset, for k = 2, …,10. For each k, the best C and γ were determined by (i) minimizing the difference between the cross-validation training accuracy and testing accuracy; (ii) maximizing the cross-validation training accuracy; and (iii) minimizing the value of C. The optimal values of C and γ were obtained by computing the Matthews correlation coefficient (MCC) for each cross-validation (k = 2, … ,10) and then choosing C and γ for which MCC is maximal, and C is minimal.

Matthews Correlation Coefficient algorithm for adjusting imbalanced datasets

The Matthews correlation coefficient (MCC) statistical test [30,31] was used to determine the optimal decision boundary threshold (for *Yes-TIL* and *No-TIL* classifications) that accounts for imbalances in the training dataset. Usually, this threshold is set to 0.5, because it is assumed that SVM works with balanced data (i.e., similar numbers of data fall into each class). For the imbalanced dataset, this threshold has to be adjusted. This adjustment was done by testing thresholds between 0.01 and 0.99 by separating the training cohort data into *Yes-TIL* or *No-TIL* classes based on whether their RBF-SVM–generated prediction probabilities exceeded the given threshold. For each threshold, MCC was calculated on the labeled prediction probabilities. The threshold with the maximum MCC was called optimal and was used as the decision boundary threshold to generate predictions on the testing cohort. In this approach, the magnitude of the decision function $\boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}_i + b$ for each $\boldsymbol{X}_i$ were extended to probability estimates using the scikit-learn library[32], with option 'probability=True'. A detailed algorithm is presented in Supplementary Figure 8.

Multiple Imputation framework for data retention

To enable the use of a dataset larger than in the complete set analysis, we implemented the Multiple Imputation with Denoising Autoencoders (MIDAS) [18,19] method. This is a deep learning technique that uses denoising autoencoders[33] and Monte Carlo dropout to generate MI with realistic uncertainty quantification. In this method, let $\boldsymbol{X} \in \mathbb{R}^{M \times P}$ denotes a data matrix with observed entries $\boldsymbol{X}_{obs}$ and missing entries $\boldsymbol{X}_{miss}$. The goal of MI is to replace each entry in $\boldsymbol{X}_{miss}$ with imputed values that preserve the interrelations in $\boldsymbol{X}_{obs}$. In our dataset 42 patients were missing data in at least one feature and the MIDAS data imputation method extended the feature dataset from 40 patients to 82 patients (leaving the blinded validation cohort of 20 patients aside). This dataset was used to identify robust predictive features and to apply the RBF-SVM and MCC algorithms. However, this approach did not improve overall predictability of our model.

Performance metrics for ML-PETIL evaluation

To assess performance of the classification protocol on the datasets that were not used for training, 2 datasets (testing cohort and a blinded validation cohort) were classified into *Yes-TIL* or *No-TIL* categories based on the parameters identified during training. The following performance metrics were used for evaluation: (i) true positive rate (TPR), or sensitivity, is the percentage of correctly classified positive instances: TPR = TP/(TP + FN); (ii) true negative rate (TNR), or specificity, is the percentage of correctly classified negative instances: TNR = TN/(FP + TN); (iii) accuracy is the percentage of correctly classified positive and negative instances: accuracy = (TP + TN)/(TP + FN + FP + TN). Here, TP (true positive) is the correctly classified *Yes-TIL* data, TN (true negative) is the correctly classified *No-TIL* data, FP (false positive) is the misclassification as *Yes-TIL*, and FN (false negative) is the misclassification as *No-TIL*.

References

1 Siegel, R. L., Giaquinto, A. N. & Jemal, A. Cancer statistics, 2024. *CA Cancer J Clin* **74**, 12-49 (2024). https://doi.org:doi:10.3322/caac.21820

2 Sylvester, R. J. *et al.* Long-term efficacy results of EORTC genito-urinary group randomized phase 3 study 30911 comparing intravesical instillations of epirubicin, bacillus Calmette-Guerin, and bacillus Calmette-Guerin plus isoniazid in patients with intermediate- and high-risk stage Ta T1 urothelial carcinoma of the bladder. *Eur Urol* **57**, 766-773 (2010). https://doi.org:10.1016/j.eururo.2009.12.024

3 Shabsigh, A. *et al.* Defining early morbidity of radical cystectomy for patients with bladder cancer using a standardized reporting methodology. *Eur Urol* **55**, 164-174 (2009). https://doi.org:10.1016/j.eururo.2008.07.031

4 Rosenberg, S. A. & Restifo, N. P. Adoptive cell transfer as personalized immunotherapy for human cancer. *Science* **348**, 62-68 (2015). https://doi.org:10.1126/science.aaa4967

5 Besser, M. J. *et al.* Clinical responses in a phase II study using adoptive transfer of short-term cultured tumor infiltration lymphocytes in metastatic melanoma patients. *Clin Cancer Res* **16**, 2646-2655 (2010). https://doi.org:10.1158/1078-0432.CCR-10-0041

6 Stevanovic, S. *et al.* Complete regression of metastatic cervical cancer after treatment with human papillomavirus-targeted tumor-infiltrating T cells. *J Clin Oncol* **33**, 1543-1550 (2015). https://doi.org:10.1200/JCO.2014.58.9093

7 Pilon-Thomas, S. *et al.* Efficacy of adoptive cell transfer of tumor-infiltrating lymphocytes after lymphopenia induction for metastatic melanoma. *J Immunother* **35**, 615-620 (2012). https://doi.org:10.1097/CJI.0b013e31826e8f5f

8 Rohaan, M. W. *et al.* Tumor-Infiltrating Lymphocyte Therapy or Ipilimumab in Advanced Melanoma. *N Engl J Med* **387**, 2113-2125 (2022). https://doi.org:10.1056/NEJMoa2210233

9 Schumacher, T. N. & Schreiber, R. D. Neoantigens in cancer immunotherapy. *Science* **348**, 69-74 (2015). https://doi.org:10.1126/science.aaa4971

10 Cancer Genome Atlas Research, N. Comprehensive molecular characterization of urothelial bladder carcinoma. *Nature* **507**, 315-322 (2014). https://doi.org:10.1038/nature12965

11 Housseau, F. *et al.* MHC-dependent cytolysis of autologous tumor cells by lymphocytes infiltrating urothelial carcinomas. *International journal of cancer* **71**, 585-594 (1997).

12 Aydin, A. M. *et al.* The Factors Affecting Expansion of Reactive Tumor Infiltrating Lymphocytes (TIL) From Bladder Cancer and Potential Therapeutic Applications. *Front Immunol* **12**, 628063 (2021). https://doi.org:10.3389/fimmu.2021.628063

13 Poch, M. *et al.* Expansion of tumor infiltrating lymphocytes (TIL) from bladder cancer. *Oncoimmunology* **7**, e1476816 (2018). https://doi.org:10.1080/2162402X.2018.1476816

14 NCT05768347. *Intravesical Adoptive Cell Therapy w/ TIL for BCG Exposed High Grade NMIBC*, <https://clinicaltrials.gov/study/NCT05768347> (2023).

15 Garapati, S. S. *et al.* Urinary bladder cancer staging in CT urography using machine learning. *Med Phys* **44**, 5814-5823 (2017). https://doi.org:10.1002/mp.12510

16 Tokuyama, N. *et al.* Prediction of non-muscle invasive bladder cancer recurrence using machine learning of quantitative nuclear features. *Mod Pathol* **35**, 533-538 (2022). https://doi.org:10.1038/s41379-021-00955-y

17 Mi, H. *et al.* Predictive models of response to neoadjuvant chemotherapy in muscle-invasive bladder cancer using nuclear morphology and tissue architecture. *Cell Rep Med* **2**, 100382 (2021). https://doi.org:10.1016/j.xcrm.2021.100382

18 Lall, R. & Robinson, T. Efficient Multiple Imputation for Diverse Data in Python and R: MIDASpy and rMIDAS. *Journal of Statistical Software* **107**, 1 - 38 (2023). https://doi.org:10.18637/jss.v107.i09

19 Lall, R. & Robinson, T. The MIDAS Touch: Accurate and Scalable Missing-Data Imputation with Deep Learning. *Political Analysis* **30**, 179-196 (2022). https://doi.org:doi:10.1017/pan.2020.49
20 Breiman, L. Random forests. *Mach Learn* **45**, 5-32 (2001). https://doi.org:Doi 10.1023/A:1010933404324
21 Svetnik, V. *et al.* Random forest: a classification and regression tool for compound classification and QSAR modeling. *J Chem Inf Comput Sci* **43**, 1947-1958 (2003). https://doi.org:10.1021/ci034160g
22 Menze, B. H. *et al.* A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinformatics* **10**, 213 (2009). https://doi.org:10.1186/1471-2105-10-213
23 Nembrini, S., Konig, I. R. & Wright, M. N. The revival of the Gini importance? *Bioinformatics* **34**, 3711-3718 (2018). https://doi.org:10.1093/bioinformatics/bty373
24 Chapelle, O., Haffner, P. & Vapnik, V. N. Support vector machines for histogram-based image classification. *IEEE Trans Neural Netw* **10**, 1055-1064 (1999). https://doi.org:10.1109/72.788646
25 Guyon, I., Weston, J., Barnhill, S. & Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach Learn* **46**, 389-422 (2002). https://doi.org:Doi 10.1023/A:1012487302797
26 Lange, A., Madiraju, S. & Petros, F. G. Therapeutic Advances in Bladder Preservation for BCG-Unresponsive Non-Muscle Invasive Bladder Cancer. *Cancers (Basel)* **17** (2025). https://doi.org:10.3390/cancers17040636
27 Balar, A. V. *et al.* Pembrolizumab monotherapy for the treatment of high-risk non-muscle-invasive bladder cancer unresponsive to BCG (KEYNOTE-057): an open-label, single-arm, multicentre, phase 2 study. *The Lancet. Oncology* **22**, 919-930 (2021). https://doi.org:10.1016/S1470-2045(21)00147-9
28 Narayan, V. M. *et al.* Efficacy of Intravesical Nadofaragene Firadenovec for Patients With Bacillus Calmette-Guerin-Unresponsive Nonmuscle-Invasive Bladder Cancer: 5-Year Follow-Up From a Phase 3 Trial. *The Journal of urology* **212**, 74-86 (2024). https://doi.org:10.1097/JU.0000000000004020
29 Lee, C. P. & Lin, C. J. A study on L2-loss (squared hinge-loss) multiclass SVM. *Neural Comput* **25**, 1302-1323 (2013). https://doi.org:10.1162/NECO_a_00434
30 Boughorbel, S., Jarray, F. & El-Anbari, M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. *PLoS One* **12**, e0177678 (2017). https://doi.org:10.1371/journal.pone.0177678
31 Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21**, 6 (2020). https://doi.org:10.1186/s12864-019-6413-7
32 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825—2830 (2011). https://doi.org:10.5555/1953048.2078195
33 Bengio, Y., Yao, L., Alain, G., and Vincent, P. Generalized denoising auto-encoders as generative models. *Proceedings of the 27th International Conference on Neural Information Processing Systems* **1**, 899–907 (2013). https://doi.org:10.5555/2999611.2999712