PETIL: Predicting Expansion of Tumor Infiltrating Lymphocytes for the Adoptive Cell Immunotherapy in Bladder Cancers

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

[*]corresponding author: Katarzyna Rejniak, address: Moffitt Cancer Center, 12902 Magnolia Dr. Tampa FL, 33612, USA, email: Kasia.Rejniak@moffitt.org

Short title: PETIL: immunotherapy stratification model for bladder tumors

Abstract

Adoptive cell therapy (ACT) with tumor-infiltrating lymphocytes (TIL) is a form of personalized immunotherapy that requires ex vivo expansion of autologous TILs and their reinfusion back into the patient. Predicting TIL expansion at the time of diagnosis may improve selection of patients that can benefit from ACT-TIL. It can also prevent high treatment-related costs and delays in treatment of patients whose cancer specimens would not yield successful TIL growth. We developed PETIL, a machine-learning model optimized for data of a medium size to determine a minimal combination of features (demographic, clinical, and biological specimen-based) that is predictive of expansion of TILs from a resected bladder cancer. We used a retrospectively identified set of data from bladder cancer patients at Moffitt Cancer Center for the training and testing cohorts. Additionally, we used data from a recent feasibility clinical trial at Moffitt Cancer Center as a blinded validation cohort. PETIL uses random forest method to identify a combination of robust predictive features, support vector machine model to determine the optimal classification hyperparameters, and Matthews correlation coefficient method to adjust the decision-boundary threshold for imbalanced data. Our model yielded AUC=0.740 for the testing cohort and AUC=0.857 for blinded validation cohort. Thus, our PETIL model optimized for data of medium size has favorable performance metrics for predicting TIL expansion from a given tumor.

Authors Summary

Treatment with autologous tumor-infiltrating lymphocytes (TIL) that are expanded ex vivo from a given tumor and then reinfused into the patient is a promising personalized immunotherapy. However, the TIL expansion takes about 4-6 weeks, thus developing tools that predict whether TIL growth will be successful can help to avoid delays in treatment of patients whose cancer specimens would not yield successful TIL expansion. Our Predictor of Expansion of TIL (PETIL) is a machine-learning model that uses patients' demographic information, clinical tumor classification, and biological tumor specimen-based measurements to determine a minimal set of these data features that are predictive of TIL expansion outcome. We applied this model to data from bladder cancer patients collected at Moffitt Cancer Center and showed that PETIL has favorable performance metrics for the dataset of a moderate size. This computational predictor can support clinicians in determining which patients are candidates for TIL immunotherapy. The developed PETIL pipeline can also be adjusted to data from other solid tumors.

Introduction

Bladder cancer is the fourth most common cancer among men and a leading cause of cancer death among men and women. The American Cancer Society estimates approximately 84,530 new cases of bladder cancer and 17,870 bladder cancer-related deaths in the United States in 2026 (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, 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). Because bladder tumors have a high mutational burden corresponding to an increased number of

neoantigens (8, 9), 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 (10), and that expanding TILs ex vivo from the resected bladder tumors is feasible (11, 12). Moreover, bladder cancer provides a unique opportunity to deliver TIL intravesically by administering T cells through a catheter into the bladder directly to tumors. An active phase I feasibility clinical trial of intravesical adoptive cell therapy with TIL for high-grade non-muscle invasive bladder cancer (NCT05768347, Moffitt Cancer Center (13)) is showing that this treatment is well-tolerated.

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 (12) 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 individual patient data. ML methods have been previously applied to several aspects of bladder cancer, such as evaluation of cancer stage in computed tomography urography (14); predictions of early recurrence of non-muscle invasive bladder cancer (NMIBC) based on histology images (15); and stratification of muscle-invasive bladder cancer (MIBC) patients into good or poor survival groups after chemotherapy (16). Here, we present the ML-based **P**redictor of the **E**xpansion of **T**umor **I**nfiltrating **L**ymphocytes (PETIL), a tool that can first learn from patient and tumor data collected in the clinic which data features are important for predicting TIL expansion, without the need to predefine which data categories to consider. Subsequently, 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-based predictor, PETIL, to stratify individual patients into *Yes-TIL* vs. *No-TIL* classes that define TIL expansion potential.

Study design

The prospective database of 106 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. The initial screening identified the maximal subset of 60 patients with complete information on 15 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, 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 46 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 -0.21884 and 0.303202 for each of the 15 features, S1 Table), the combinations of features were considered, and the ML approach was applied. For every patient, information was recorded on whether the resected tumor sample caused positive TIL growth. Among the collected datasets, 68 tumors demonstrated TIL expansion (class *Yes-TIL*), and 38 did not (class *No-TIL*). This information was used for training and testing the PETIL predictor.

Table 1: Description of Study Population

Features	Retrospective Dataset		Clinical Trial Dataset	
	Total Cohort (N=106)	**Missing data, n**	**Total Cohort (N=14)**	**Missing data, n**
Age at Surgery, mean ± SD	69.3 ± 11.6	0	70.4 ± 6.9	0
Race, n (%)		0		0
American Indian/Alaskan Native	1 (0.9)		0	
Asian	1 (0.9)		0	
Black	2 (1.9)		0	
White	101 (95.3)		14 (100.0)	
Unknown	1 (0.9)		0	
Surgery, n (%)		0		0
Radical Cystectomy	73 (68.9)		4 (28.6)	
TURBT	33 (31.1)		10 (71.4)	
Smoker, n (%)		0		0
Current Smoker	15 (14.2)		1 (7.1)	
Ever Smoked	58 (54.7)		7 (50.0)	
Never Smoked	23 (21.7)		6 (42.9)	
Not Reported	10 (9.4)		0	
BMI, mean ± SD	29.3 ± 7.6	4	28.7 ± 5.8	0
NAC, n (%)		3		9
NAC	41 (39.8)		3 (60.0)	
No NAC	62 (60.2)		2 (40.0)	
cT, n (%)		9		0
Ta	18 (18.6)		5 (35.7)	
CIS	2 (2.1)		1 (7.1)	
T1	17 (17.5)		5 (35.7)	
T2	41 (42.3)		3 (21.4)	
T3	14 (14.4)		0	
T4	5 (5.2)		0	
pT, n (%)		21		10
T0	5 (5.9)		0	
Ta	8 (9.4)		1 (25.0)	
CIS	3 (3.5)		1 (25.0)	
T1	5 (5.9)		0	
T2	20 (23.5)		1 (25.0)	
T3	29 (34.1)		0	
T4	15 (17.6)		1 (25.0)	
pN, n (%)		33		10
N0	51 (69.9)		4 (100.0)	
N1	9 (12.3)		0	
N2	5 (6.8)		0	
N3	6 (8.2)		0	
Nx	2 (2.7)		0	

Bx Histology, n (%)		2		9
Benign	2 (1.9)		0	
Urothelial Carcinoma	99 (95.2)		5 (100.0)	
Squamous Cell Carcinoma	2 (1.9)		0	
Small Cell Carcinoma	1 (1.0)		0	
Histology, n (%)		3		10
Benign	2 (1.9)		0	
Urothelial Carcinoma	99 (96.1)		4 (100.0)	
Squamous Cell Carcinoma	2 (1.9)		0	
Sample weight (g), mean $\pm$ SD	1.2 ± 2.3	0	2.7 ± 8.5	0
Tumor digest count, mean $\pm$ SD	$1.43 \pm 3.38 \times 10^7$	0	$6.77 \pm 14.83 \times 10^7$	9
Fragments plated, n (%)		0		0
0 – 5	21 (19.8)		1 (7.1)	
6 – 11	22 (20.8)		6 (42.9)	
12 – 17	19 (17.9)		4 (28.6)	
18 – 23	8 (7.5)		0	
>= 24	36 (34.0)		3 (21.4)	

Abbreviations: 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.

The PETIL pipeline shown in Figure 1 consists of four steps: (i) data pre-processing that includes splitting data 70/30 into training and testing cohorts, data values normalization and imputation of the missing data, removal of features that are correlated, and analysis of ML classification methods for which the data sample size is adequate to drawn predictive conclusions; (ii) selection of a smaller subset of features that are robust in making predictions; (iii) model training that includes learning hyperparameters for the chosen ML predictor and determining the decision boundary threshold to account for imbalanced data; and (iv) generating predictions and calculating the performance metrics. After training, PETIL has established a minimal list of predictive features, a ML classifier adequate to the given data, and the decision threshold for the imbalanced data. Subsequently, PETIL was applied to stratify patients in the testing cohort. Finally, the performance metrics were computed, including accuracy, sensitivity, and specificity.

For additional validation, the same type of data was collected from the phase I clinical trial "Intravesical Adoptive Cell Therapy w/ TIL for BCG Exposed High Grade NMIBC", that was conducted at Moffitt Cancer Center between 2023 and 2025 (NCT05768347, (13)) under the protocol approved by the Advarra Institutional Review Board (MCC21894). Informed consent was also obtained from all patients prior to data collection. The summary of this data is also shown in Table 1. However, the corresponding TIL growth status was blinded until after the PETIL predictor was developed and the predictions for this cohort were generated. This dataset was also subject to data normalization and imputation of the missing data before PETIL generated predictions. Even if the clinical trial dataset was small, it was used as a blinded validation cohort due to the lack of a validation dataset from another institution, as the dataset we used here is the biggest collection of bladder cancer TIL specimens in one institution.

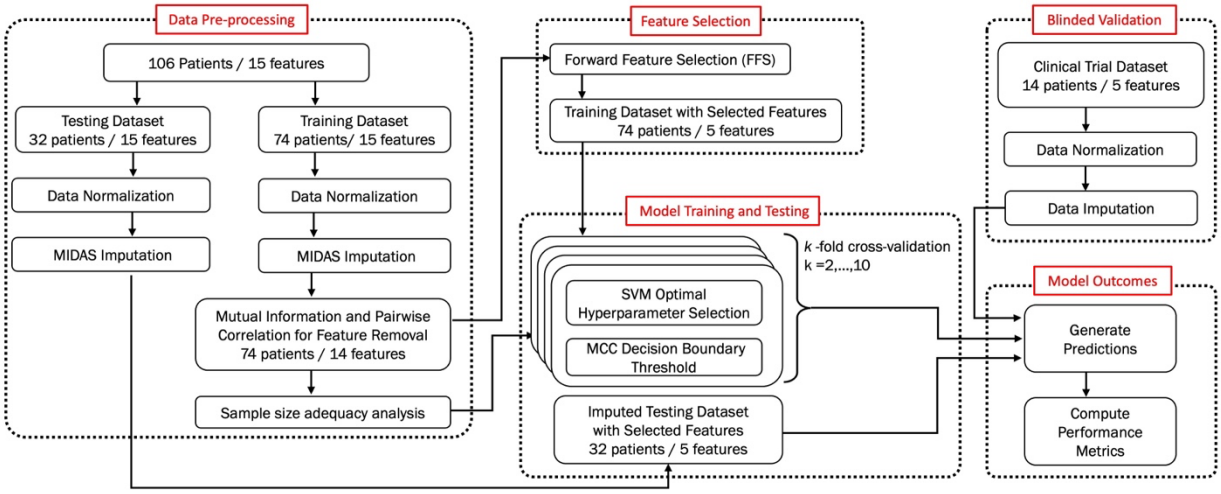


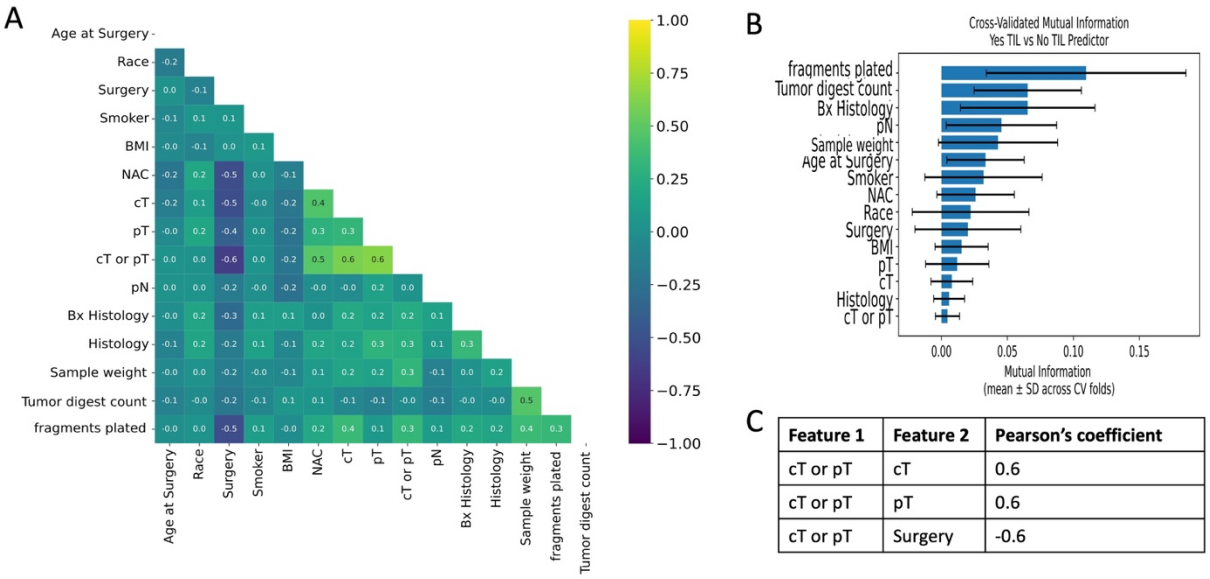
Figure 1. The PETIL pipeline. Abbreviation: PETIL: Predictor of the Expansion of Tumor Infiltrating Lymphocytes; MIDAS: Multiple Imputation with Denoising Autoencoders; FFS: Forward Feature Selection; SVM: Support Vector Machines.

Data normalization and imputation

After splitting 70/30 the retrospective dataset into training (74 patients) and testing (32 patients) cohorts, both were normalized separately using *MaxAbsScaler,* a class in the Scikit-learn Python library (17). The total proportion of missingness in the overall dataset was 4.4%, with 7 features having between 2 and 33 missing entries (Table 1). To address the issue of missing entries, we employed the Multiple Imputation with Denoising Autoencoders (MIDAS (18)) method, a deep learning approach that can generate imputed values while preserving interrelations among all features. MIDAS was applied to the training and testing cohorts separately and learned and imputed admissible values for the missing data. For the clinical cohort used for blinded validation, the data was normalized and all missing entries were imputed with zero value due to the small size (n = 14) of this cohort.

Identification of pairwise collinear features

Using the Pearson's pairwise correlation coefficient and setting the collinear threshold to 0.6 or higher (high positive correlation) and -0.6 or lower (high negative correlation), we tested all 15 features in the imputed training dataset to identify collinear features (Figure 2A). Next, we ranked the 15 features according to their Mutual Information (MI (19)) scores (Figure 2B). The Feature 'cT or pT' was the only one that met the collinearity threshold with either 'pT', 'cT', or 'Surgery' features, but due to lower MI score for 'cT or pT', this feature was removed from further consideration (Figure 2C). This reduced the dimension of the feature space from 15 to 14.



C

Feature 1	Feature 2	Pearson's coefficient
cT or pT	cT	0.6
cT or pT	pT	0.6
cT or pT	Surgery	-0.6

Figure 2. Identification of collinear features. A. Pearson's Pairwise correlation for all 15 features. **B.** Mutual Information score for all 15 features **C.** The collinear features from **A** that met the cut-off threshold, due a lower MI score in **B** feature 1: 'cT or pT' will be removed**.**

.

Sample size adequacy analysis

We conducted a sample size adequacy assessment by using the learning curves analysis (20, 21) to identify an appropriate ML classifier for the given task and given data sample. We showed that for the training dataset of size 74, the support vector machines (22-24) with the radial basis function kernel (RBF-SVM) were either optimal or performed better than the logistic regression (25), gradient boosting (26), and random forest (27) classifiers. The learning curves were obtained by stratified k-fold cross-validation on subsets of the training dataset containing from 10% to 100% of the available training dataset and evaluating performance on the held-out validation dataset using the RBF-SVM classifier (Figure 3). The learning curves show high validation scores for most of the sampled subsets of the training dataset. The low variance between the training scores and the validation scores, as well as the near convergence in the validation scores, indicates high generalization to new data and adequacy of the 74-training dataset for classification task. The learning curve analysis for the remaining three classifiers is shown in S1 Figure.

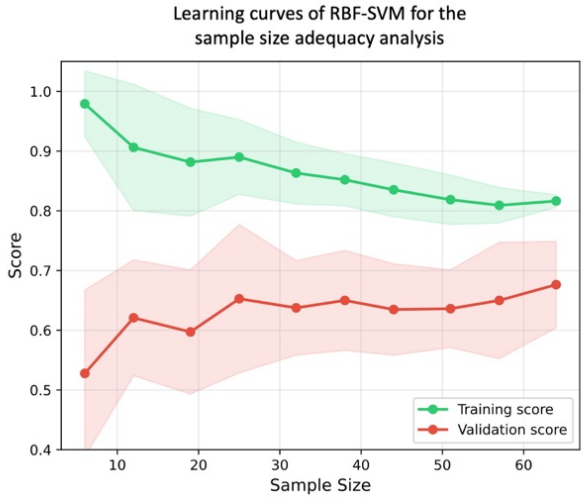


Figure 3. Sample size adequacy analysis for the RBF-SVM predictor. Learning curves show training (green) and validation (red) performance across training datasets of varying sample sizes. Shaded regions represent standard deviation across 8-fold cross-validation. The validation curve low variance with the training curve indicates that the classifier does not overfit to the 74-training set with 14 features. The validation curve show that performance analysis could be improved with more training data.

Identification of a combination of robust predictive features

The 14 features commonly collected in the clinic were used as a base for identifying a smaller set of robust predictive features. The base set included 4 demographic, 7 clinical, and 3 biological tumor specimen-based features, all listed in the Materials and Methods section. Using the training dataset, we implemented the forward feature selection (FFS) method (28) to identify a subset of features that can distinguish between the *Yes-TIL* and *No-TIL* classes. FFS determined a pre-ranking of the 14 features by calculating the feature importance score (FIS) (27, 29, 30) using the random forest (RF) classifier on a stratified cross-validated 10 subsamplings of the training dataset (Figure 4A). Next, we used each fold of the 10 stratified cross-validated subsamples of the training dataset to train the RF classifier, where we sequentially added each feature in the order of its ranking and validated the addition of each feature on the holdout in each fold. This process of adding features one at a time yields a non-strictly increasing curve of the RF accuracy on the holdout sets. There may be local regions of downward fluctuations due to noisy features. To mitigate these fluctuations, the validation accuracy curve was converted to a monotonically increasing curve, retaining the maximal accuracy observed up to each feature addition step. The optimal number of features was determined using a stopping criterion that checks for plateau in the validation accuracy curve (31), beyond which additional features provide minimal incremental gain in accuracy (Figure 4B).

The FFS algorithm determined 5 robust predictive features: (i) two demographic features (age at surgery, BMI) and (ii) three biological tumor specimen-based features (number of fragments plated, sample weight tumor, tumor digest count) shown boxed in Figure 4A. The distributions of the robust predictive feature values in the training and testing cohorts are shown in S2 Figure.

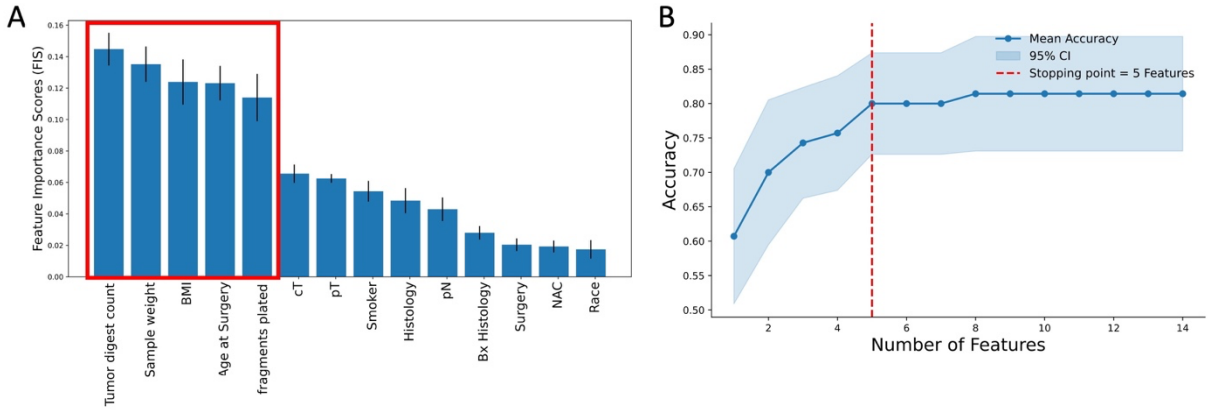


Figure 4. Feature selection. A. The 14 features were pre-ranked using the FIS values obtained from training the RF algorithm on a stratified cross-validated subsamples of the training dataset. **B.** The FFS algorithm calculated the validation accuracy of sequentially updated features with high FIS rank and then using a stopping criterion, it identified 5 robust predictive features.

Learning of the predictive SVM and MCC hyperparameters

The PETIL classifier of choice is the RBF-SVM because the learning curve analysis indicated its superior data generalization and adequacy of the size of the training dataset for classification in comparison with other classifier, such as logistic regression, gradient boosting, and random forest. Another advantage of the RBF-SVM classifier is its ability to capture interactions between predictive features that are potentially multidimensional and nonlinear (23, 24). Using the training dataset with the selected 5 robust predictive features, we learned the optimal classification hyperparameters using a *k*-fold cross-validation (for k = 2, …,10). For each *k,* we used the Matthews correlation coefficient (MCC) to determine the best decision boundary threshold so as to account for the imbalance in the dataset. For each *k,* we identified the best RBF-SVM hyperparameters (C, γ). Here, C specifies the width of the margins for avoiding data misclassification, while γ controls the nonlinearity of the decision boundary hyperplane. The best C and γ were determined by (i) minimizing the difference between the cross-validation training accuracy and holdout accuracy; (ii) maximizing the cross-validation training accuracy; and (iii) minimizing the value of C. We determined the optimal *k=5,* and the corresponding optimal hyperparameters values were C = 2.520 and γ = 0.02 (S3 Figure). The optimal MCC decision boundary threshold was 0.629.

Evaluating the predictive performance of PETIL

The developed optimal RBF-SVM-MCC model was first verified using the testing dataset and was subsequently validated using the blinded validation (clinical trial dataset). The testing cohort comprised 32 patients, and the PETIL predictions yielded the accuracy of 0.750, the true positive rate (sensitivity) of 0.762, and the true negative rate (specificity) of 0.727 (Figure 5A). The area under the curve (AUC) for the testing cohort was 0.740 (Figure 5B). The clinical trial dataset comprised 14 patients and all successfully had TIL expansion. PETIL predicted this outcome with an accuracy of 0.857 (12 out of 14 correct predictions).

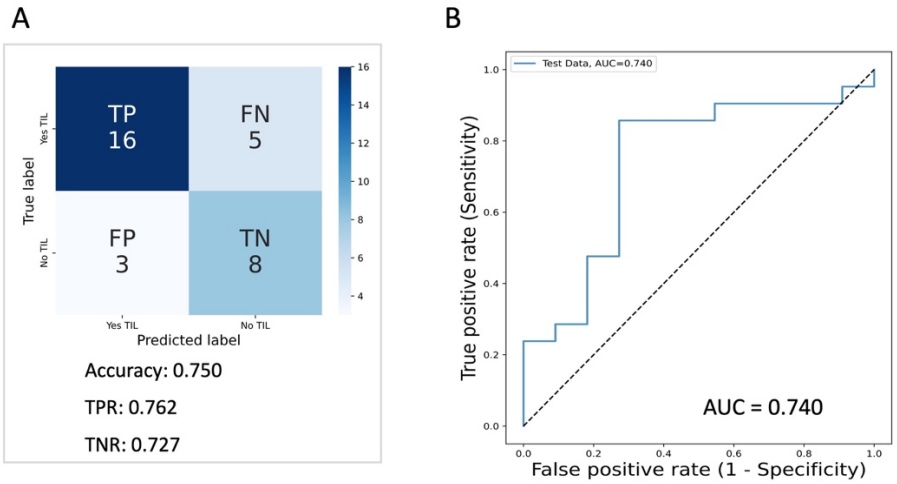


Figure 5 Model performance analysis. A. The confusion matrix of PETIL predictions for the testing dataset with the indicated performance metrics: the accuracy, true positive rate (TPR), and true negative rate (TNR). **B.** The receiver-operating characteristic (ROC) curve generated for PETIL predictions for the testing dataset with the indicated area under the curve (AUC). The dashed line represents a random classifier. Abbreviations: PETIL, Predictor of the Expansion of Tumor Infiltrating Lymphocytes.

Discussion

Despite recent advancements in treatment of non-muscle invasive bladder cancer (32), 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 (33, 34). 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 (12). Moreover, *ex vivo* expansion of TIL takes about 4 weeks after tumor resection, thus early assessment of the outcome of TIL expansion could benefit the patient. Early identification of patients who could benefit from ACT-TIL is essential for optimizing treatment strategies and in preventing delays in next appropriate treatment. For example, 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.

Here, we conducted a retrospective analysis of multimodal data of 106 bladder cancer patient specimens and developed the 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 PETIL predictor showed favorable performance metrics on both the testing cohort that was a part of the retrospective database (AUC = 0.740) and the validation cohort that came from an active clinical trial (Accuracy of 0.857, correctly predicted 12 out of 14 TIL expansions).

Another important feature of the PETIL classifier is the capacity to make successful predictions using cohorts of a medium size (74 datasets for a training cohort and 32 for a testing cohort). This is in contrast to other ML methods, and deep learning methods in particular, that require large datasets for training. A limitation in our study is the lack of an external validation dataset from another institution. However, we utilized the phase I clinical trial dataset (NCT05768347) as a blinded validation cohort. To our knowledge, the combined dataset we used (11, 12) 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.

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. This model was able to predict TIL expansion with favorable performance. While further studies with larger cohorts of data are still desirable, PETIL predictor shows a great promise to be used as a clinical-supportive tool to help with stratifying patient eligibility for immune-based therapies, such as the ACT-TIL.

Materials and Methods

Clinical and experimental data

The prospective database of 106 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. Additional data for 14 adult patients was collected from the phase I clinical trial (NCT05768347) conducted at Moffitt Cancer Center between 2023 and 2025 under the protocol approved by the Advarra Institutional Review Board (MCC21894). Informed consent was also 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 (12). 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*. All other cases were categorized as *No-TIL*. This established database that was used for retrospective data analysis and ML-based predictions.

All datasets included information on 15 commonly collected features about patients' (i) demographics: 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 (pN), type of surgery, prior neoadjuvant chemotherapy (NAC), 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 tumor digest count. In total, 56 patients were missing data in some considered features. We used the MIDAS imputation

method, since it can handle both numerical and categorical data features. For every patient, information was recorded on whether the resected tumor sample caused positive TIL growth. For the clinical trial validation cohort, the TIL growth status was blinded until after the PETIL predictor was developed and the predictions for this cohort generated. The clinical trial dataset was used as a validation cohort due to the lack of a validation dataset from another institution, since the dataset we used here is the biggest collection of bladder cancer TIL specimens in one institution.

Description of the ML classification problem

Our goal was to identify a predictive subset of features that differentiates between two targets in a binary classification task, and to provide metrics of success for such data stratification. Let the dataset $\mathbf{X}$ consists of M data points: $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_\mathrm{M}]^\intercal$, where each data point has P features: $\mathbf{X}_\mathrm{i} = \left(\mathbf{x}_\mathrm{i}^1, \mathbf{x}_\mathrm{i}^2, \cdots, \mathbf{x}_\mathrm{i}^\mathrm{P}\right)$ for $\mathrm{i} \in \{1,2,\cdots,\mathrm{M}\}$, and for each i, the corresponding target is the binary class $\mathrm{y}_\mathrm{i} \in \{-1, 1\}$. The goal was to find the minimal subset of features of size Q, where $\mathrm{Q} < \mathrm{P}$, that divides the dataset $\mathbf{X}$ into distinct binary classes. We used a data-informed approach. First, we split $\mathbf{X}$ into training and testing cohorts. Using the training cohort, we implemented feature selection method to identify predictive features. Next, we determined an optimal machine learning classifier using the sample size adequacy methods. Subsequently, we found the optimal hyperparameters for the classifier using a cross-validation technique and also learned the optimal decision boundary threshold that corrects for imbalance in the binary classes in $\mathbf{X}$. Finally, this classifier was applied to the testing cohort to assess the prediction metrics.

Data normalization

Data normalization is performed to ensure that all features can contribute equally. In order to prevent data leakage between training and testing cohorts, we split the overall data before data normalization, and perform normalization on the training and testing cohorts separately. We used the *MaxAbsScaler*, a class in the scikit-learn Python library (17) which translates each feature independently to have a maximal absolute value of 1.0, preserving data distribution but only linearly scales down each feature. Data normalization was also used for the clinical trial cohort.

Multiple Imputation framework for data retention

Multiple imputation is a statistical method to handle missing data. Let $\mathbf{X} \in \mathbb{R}^{\mathrm{M}\times\mathrm{P}}$ be a data matrix with observed entries $\mathbf{X}_{\mathrm{obs}}$ and missing entries $\mathbf{X}_{\mathrm{miss}}$. Under the assumption that data are missing at random (MAR) or completely at random (MCAR), the multiple imputation replaces all entries in $\mathbf{X}_{\mathrm{miss}}$ with imputed values that preserve the interrelations in $\mathbf{X}_{\mathrm{obs}}$. We use here the multiple imputation with denoising autoencoders (MIDAS) method (18), which is a scalable deep learning-based technique that employs a class of unsupervised neural networks known as denoising autoencoders (35) and Monte Carlo dropout to generate multiple imputation of the missing data with realistic uncertainty quantification. In our training dataset 32 out of 74 patients were missing data in at least one feature in the training dataset, and 14 out of 32 patients were missing data in at least one feature in the testing dataset. In either cohort, the MIDAS data imputation method was used to impute missing data. The imputed training dataset was used to identify robust predictive features and to apply the RBF-SVM and MCC algorithms.

Mutual information measure for nonlinear dependencies

Mutual information (MI) is a non-parametric measure of statistical dependency between the dataset $\mathbf{X} \in \mathbb{R}^{\mathrm{M}\times\mathrm{P}}$ and the predicted binary class $\mathbf{Y} \in \mathbb{R}^{\mathrm{M}\times 1}$, where M is the number of patients and P is the number of features. MI captures nonlinear dependencies in high-dimensional data, which makes them robust for measuring feature relevance in discrete or categorical datasets (36, 37). A discrete MI is computed as follows: $\mathrm{MI}(\mathbf{X}, \mathbf{Y}) = \sum_{\mathrm{x}\in\mathbf{X}}\sum_{\mathrm{y}\in\mathbf{Y}} \mathrm{p}(\mathrm{x},\mathrm{y})\ \log\frac{\mathrm{p}(\mathrm{x},\mathrm{y})}{\mathrm{p}(\mathrm{x})\,\mathrm{p}(\mathrm{y})}$. Where $\mathrm{p}(\mathrm{x},\mathrm{y})$

is the joint probability of $\mathbf{X}$ and $\mathbf{Y}$, while $\mathrm{p}(\mathrm{x})$ and $\mathrm{p}(\mathrm{y})$ are the marginal probabilities of $\mathbf{X}$ and $\mathbf{Y}$, respectively.

Learning curve method for determining sample size adequacy

To assess training dataset adequacy for a classification task of distinguishing TIL expansion status, we performed learning curve analysis using stratified k-fold cross-validation, where we repeatedly subsampled the training data at different sample sizes. We trained four models (Logistics Regression, Random Forest, Gradient Boosting, and RBF-SVM) on subsets of the training dataset from 10% to 100% of the available training dataset and evaluated performance on the holdout validation dataset.

Forward feature selection method for identification of predictive features

The feature selection was performed to identify the minimal predictive subset of features used in the data classifier. Using the Forward Feature Selection (FFS) method (28), we built a stratified cross-validated 10 subsamplings of the training dataset. On each fold, we trained a random forest (RF) classifier and obtained feature importance score (FIS) ranking for each of the features. Next, the features were added sequentially in the order of pre-ranking, and the predictive accuracy of the RF classifier was evaluated using the holdout validation set. To mitigate fluctuations due to noisy features, accuracy curves were converted to a monotonic increasing curve, retaining the maximum accuracy observed up to each feature addition step. The optimal number of features for each fold was determined using a stopping criterion which identifies the point beyond which additional features provide minimal incremental gain in accuracy. This stopping criterion also prevents overfitting to the training data. The mean accuracy curve and 95% confidence interval across all folds were used to visualize feature contribution to the predictiveness of the model and to determine final feature selection.

Support vector machine model for learning the optimal classification rules

For each data point $\mathbf{X}_\mathrm{i} = \left(\mathbf{x}_\mathrm{i}^1, \mathbf{x}_\mathrm{i}^2, \cdots, \mathbf{x}_\mathrm{i}^\mathrm{P}\right)$ in $\mathbf{X} \in \mathbb{R}^{\mathrm{M}\times\mathrm{P}}$, a binary ML classifier learns a corresponding prediction y_i in $\mathbf{Y} \in \mathbb{R}^{\mathrm{M}\times 1}$. The accuracy of this prediction depends on how well the classifier identifies an optimal decision boundary That separates the $\mathbf{X}$ into the distinct binary classes. However, the interactions between the predictive features are often multidimensional and nonlinear. A support vector machine (SVM) model with the nonlinear Radial Basis Function (RBF) kernel: $\mathrm{K}\left(\mathbf{X}_\mathrm{i}, \mathbf{X}_\mathrm{j}\right) = \exp\left(-\gamma\left\|\mathbf{X}_\mathrm{i} - \mathbf{X}_\mathrm{j}\right\|^2\right)$ (23, 38) can potentially capture this complex interaction between features as it learns to distinguish between the distinct classes in the training dataset and generalize to new dataset. First, all data was split 70/30 into the training and testing cohorts. Then, the optimal RBF-SVM hyperparameters (C, γ) were identified by performing k-fold cross-validation (CV) on the training dataset, for k = 2, …,10. Here C specifies the width of the margins for avoiding data misclassification and 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 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 (39, 40) 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 $\mathbf{w}^{\mathrm{T}}\mathbf{X}_{\mathrm{i}} + \mathrm{b}$ for each $\mathbf{X}_{\mathrm{i}}$ were extended to probability estimates using the scikit-learn library (41), with option 'probability=True'. A detailed algorithm is presented in S1 Algorithm.

Performance metrics for PETIL evaluation

To assess performance of the classification protocol on the testing dataset, the following performance metrics were used for evaluation: (1) true positive rate (TPR) or sensitivity, is the percentage of correctly classified positive instances: TPR=TP/(TP+FN); (2) true negative rate (TNR) or specificity, is the percentage of correctly classified negative instances: TNR=TN/(FP+TN); (3) accuracy is the percentage of correctly classified positive and negative instances: accuracy=(TP+TN)/(TP+FN+FP+FN); (4) area under the receiver operating characteristics curve (AUC/ROC or AUC) measures the ability to discriminate between positive and negative cases and ranges from 0.5 (coin toss) to 1.0 (perfect classification), when ROC curve shows tradeoffs between TP and FP. Here, TP (true positive) is the correctly classified data, TN (true negative) is the correctly classified data, FP (false positive FP) is the misclassification of the positive class, and FN (false negative) is the misclassification of the negative class.

Supporting Information Captions

S1 Table: Pearson Correlation Coefficient between each of 15 individual data features and the TIL growth status

S1 Figure: **Learning curve analysis for different ML classifiers. A.** Learning curves for the logistic regression (LR) model. Low validation scores indicate that LR is not an adequate classifier. **C.** Learning curves for the random forest (RF) method. High variance between the training and validation scores shows that RF is not an adequate classifier. **D.** Learning curves for the gradient boosting (GB) method. High variance between the training and validation scores shows that GB is not an adequate classifier. In each analysis, sample proportions ranging from 10% to 100% of the training dataset were sampled multiple times. A stratified k-fold cross-validation was used to obtain the training (shown in green) and validation (shown in red) performance curves across varying training dataset sample sizes. Shaded regions represent standard deviation across 8-fold cross-validation.

S2 Figure: Distributions of robust predictive features. Distributions of the *No-TIL* (blue) and *Yes-TIL* (orange) classes for the training (left) and testing (right) cohorts for five robust predictive features: **A.** Tumor sample weight, **B.** Tumor digest count, **C.** Age at surgery, **D.** BMI, **E.** Number of fragments plated.

S3 Figure: Optimal hyperparameter search. Hyperparameter space for parameters C and γ for the RBF-SVM with cross-validation k=5, for testing (left) and training (right) cohorts.

S1 Algorithm: Algorithm for determining the optimal decision threshold for radial basis-kernel function for the support vector machine (RBF-SVM) and the Matthews correlation coefficient (MCC) methods.

Acknowledgments

Assistance in data collection and de-identification was provided by Ms. Suzanne McFarland, an honest broker at Moffitt Cancer Center.

Ethics Statement

These studies were reviewed and approved by the Moffitt Scientific Review Committee (SRC) and Advarra Institutional Review Board (MCC18142, MCC20106, and MCC21894) and performed in accordance with IRB guidelines. Informed consent was obtained from all patients prior to data collection.

Code availability

Computational code used in this study is available from the following GitHub depositories:
https://github.com/okayode/Predictor_of_the_Expansion_of_TIL_project,
https://github.com/rejniaklab/Predictor_of_the_Expansion_of_TIL_project

Data availability

All data used in this study is available from the following GitHub links:
https://github.com/okayode/Predictor_of_the_Expansion_of_TIL_project,
https://github.com/rejniaklab/Predictor_of_the_Expansion_of_TIL_project

Author Contributions

K.D.O., A.M.A., I.C., S.P.T., M.A.P., and K.A.R. participated in study design. A.M.A., S.B., B.B., S.P.T., and M.A.P. were responsible for data collection. K.D.O., A.K., S.P.T., M.A.P., and K.A.R. developed the methodology. K.D.O. and K.A.R. performed the formal analysis and wrote the initial draft of the manuscript, which was critically reviewed by A.M.A, A.K., S.P.T., and M.A.P. All authors approved the final version of the manuscript.

Competing Interests

Moffitt Cancer Center has licensed Intellectual Property (IP) related to the proliferation and expansion of tumor infiltrating lymphocytes (TILs) to Iovance Biotherapeutics. Moffitt has also licensed IP to Tuhura Biopharma. Dr. Pilon-Thomas (SPT) is an inventor on such Intellectual Property. SPT is listed as a co-inventor on a patent application with Provectus Biopharmaceuticals. SPT participates in sponsored research agreements with Provectus Biopharmaceuticals, Celgene, Iovance Biotherapeutics, Intellia Therapeutics, Dyve Biosciences, and Turnstone Biologics. SPT has received consulting fees from Seagen Inc., Morphogenesis, Inc., Iovance Biotherapeutics, and KSQ Therapeutics. Other authors declare no financial competing interests.

Funding

This work was supported by the Department of Defense grants W81XWH-22-1-0339, W81XWH-22-1-0340, and W81XWH-22-1-0341 (to MP, KR, and SPT), and the US National Institutes of Health, National Cancer Institute grant R01-CA259387 (to KR and SPT). This work was supported in part by the Shared Resources at the H. Lee Moffitt Cancer Center & Research Institute an NCI designated Comprehensive Cancer Center under the grant P30-CA076292 from the National Institutes of Health. The funders played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

References

1. Key Statistics for Bladder Cancer. The American Cancer Society The American Cancer Society; 2026
2. Sylvester RJ, Brausi MA, Kirkels WJ, Hoeltl W, Calais Da Silva F, Powell PH, 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. 2010;57(5):766–73.
3. Shabsigh A, Korets R, Vora KC, Brooks CM, Cronin AM, Savage C, et al. Defining early morbidity of radical cystectomy for patients with bladder cancer using a standardized reporting methodology. Eur Urol. 2009;55(1):164–74.
4. Rosenberg SA, Restifo NP. Adoptive cell transfer as personalized immunotherapy for human cancer. Science. 2015;348(6230):62–8.
5. Besser MJ, Shapira-Frommer R, Treves AJ, Zippel D, Itzhaki O, Hershkovitz L, 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. 2010;16(9):2646–55.
6. Stevanovic S, Draper LM, Langhan MM, Campbell TE, Kwong ML, Wunderlich JR, et al. Complete regression of metastatic cervical cancer after treatment with human papillomavirus-targeted tumor-infiltrating T cells. J Clin Oncol. 2015;33(14):1543–50.
7. Pilon-Thomas S, Kuhn L, Ellwanger S, Janssen W, Royster E, Marzban S, et al. Efficacy of adoptive cell transfer of tumor-infiltrating lymphocytes after lymphopenia induction for metastatic melanoma. J Immunother. 2012;35(8):615–20.
8. Schumacher TN, Schreiber RD. Neoantigens in cancer immunotherapy. Science. 2015;348(6230):69–74.
9. Cancer Genome Atlas Research N. Comprehensive molecular characterization of urothelial bladder carcinoma. Nature. 2014;507(7492):315–22.
10. Housseau F, Zeliszewski D, Roy M, Paradis V, Richon S, Ricour A, et al. MHC-dependent cytolysis of autologous tumor cells by lymphocytes infiltrating urothelial carcinomas. International journal of cancer. 1997;71(4):585–94.
11. Aydin AM, Bunch BL, Beatty M, Hajiran A, Dhillon J, Sarnaik AA, et al. The Factors Affecting Expansion of Reactive Tumor Infiltrating Lymphocytes (TIL) From Bladder Cancer and Potential Therapeutic Applications. Front Immunol. 2021;12:628063.
12. Poch M, Hall M, Joerger A, Kodumudi K, Beatty M, Innamarato PP, et al. Expansion of tumor infiltrating lymphocytes (TIL) from bladder cancer. Oncoimmunology. 2018;7(9):e1476816.
13. NCT05768347. Intravesical Adoptive Cell Therapy w/ TIL for BCG Exposed High Grade NMIBC 2023 [Available from: https://clinicaltrials.gov/study/NCT05768347].
14. Garapati SS, Hadjiiski L, Cha KH, Chan HP, Caoili EM, Cohan RH, et al. Urinary bladder cancer staging in CT urography using machine learning. Med Phys. 2017;44(11):5814–23.
15. Tokuyama N, Saito A, Muraoka R, Matsubara S, Hashimoto T, Satake N, et al. Prediction of non-muscle invasive bladder cancer recurrence using machine learning of quantitative nuclear features. Mod Pathol. 2022;35(4):533–8.
16. Mi H, Bivalacqua TJ, Kates M, Seiler R, Black PC, Popel AS, et al. Predictive models of response to neoadjuvant chemotherapy in muscle-invasive bladder cancer using nuclear morphology and tissue architecture. Cell Rep Med. 2021;2(9):100382.
17. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–30.
18. Lall R, Robinson T. Efficient Multiple Imputation for Diverse Data in Python and R: MIDASpy and rMIDAS. J Stat Softw. 2023;107(9):1–38.
19. Sun L, Xu J. Feature selection using mutual information based uncertainty measures for tumor classification. Biomed Mater Eng. 2014;24(1):763–70.

20. Perlich C, Provost F, Simonoff JS. Tree induction vs. logistic regression: a learning-curve analysis Journal of Machine Learning Research. 2003;4:211—55.
21. Figueroa RL, Zeng-Treitler Q, Kandula S, Ngo LH. Predicting sample size required for classification performance. BMC Med Inform Decis Mak. 2012;12:8.
22. Ben-Hur A, Ong CS, Sonnenburg S, Scholkopf B, Ratsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008;4(10):e1000173.
23. Chapelle O, Haffner P, Vapnik VN. Support vector machines for histogram-based image classification. IEEE Trans Neural Netw. 1999;10(5):1055–64.
24. Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector machines. Mach Learn. 2002;46(1-3):389–422.
25. Tolles J, Meurer WJ. Logistic Regression: Relating Patient Characteristics to Outcomes. JAMA. 2016;316(5):533–4.
26. Sagi O, Rokach L. Approximating XGBoost with an interpretable decision tree. Inform Sciences. 2021;572:522–42.
27. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
28. Borboudakis G, Tsamardinos I. Forward-Backward Selection with Early Dropping. Journal of Machine Learning Research. 2019;20.
29. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, 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. 2009;10:213.
30. Nembrini S, Konig IR, Wright MN. The revival of the Gini importance? Bioinformatics. 2018;34(21):3711–8.
31. Yu S, Principe JC. Simple Stopping Criteria for Information Theoretic Feature Selection. Entropy (Basel). 2019;21(1).
32. Lange A, Madiraju S, Petros FG. Therapeutic Advances in Bladder Preservation for BCG-Unresponsive Non-Muscle Invasive Bladder Cancer. Cancers (Basel). 2025;17(4).
33. Balar AV, Kamat AM, Kulkarni GS, Uchio EM, Boormans JL, Roumiguie M, 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. Lancet Oncol. 2021;22(7):919–30.
34. Narayan VM, Boorjian SA, Alemozaffar M, Konety BR, Shore ND, Gomella LG, 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. J Urol. 2024;212(1):74–86.
35. 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. 2013;1:899–907.
36. Kraskov A, Stogbauer H, Grassberger P. Estimating mutual information. Phys Rev E Stat Nonlin Soft Matter Phys. 2004;69(6 Pt 2):066138.
37. Loganathan G, Palanivelan M. Ovarian cancer detection from mutual information-ranked clinical biomarkers using an explainable attention-based residual multilayer perceptron. Comput Biol Chem. 2025;120(Pt 2):108714.
38. Lee CP, Lin CJ. A study on L2-loss (squared hinge-loss) multiclass SVM. Neural Comput. 2013;25(5):1302–23.
39. Boughorbel S, Jarray F, El-Anbari M. Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric. PLoS One. 2017;12(6):e0177678.
40. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):6.
41. Pedregosa F, Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825—30.