Am J Transl Res 2026;18(5):4153-4172
www.ajtr.org /ISSN:1943-8141/AJTR0171605

Original Article

Development and validation of a multidimensional machine-learning model to predict clinically significant immune-related adverse events in advanced non-small cell lung cancer

Xiaoqin Weng[1], Hengrui Zhang[1], Mao Huang[2]

[1]*Department of Respiratory and Critical Care Medicine, The Second Affiliated Hospital of Nanjing Medical University, Nanjing 210000, Jiangsu, China;* [2]*Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital with Nanjing Medical University, Nanjing 210000, Jiangsu, China*

Received January 21, 2026; Accepted April 7, 2026; Epub May 15, 2026; Published May 30, 2026

Abstract: Objectives: To overcome the limitations of existing predictors for immune-related adverse events (irAEs) in advanced non-small cell lung cancer (NSCLC), including failure to account for competing mortality risks and non-linear interactions, we aimed to develop an accurate machine learning model for clinically significant irAEs (cs-irAEs, Grade ≥ 2). Methods: We enrolled 332 patients with stage IIIB-IV NSCLC treated with PD-1/PD-L1 inhibitors, and randomly assigned to training (n = 232) and testing (n = 100) sets. The Fine-Gray model, adjusting for death as a competing event, estimated cs-irAE incidence. Least Absolute Shrinkage and Selection Operator (LASSO) regression selected predictors, followed by an Extreme Gradient Boosting (XGBoost) model compared to logistic regression. Model performance was assessed using under the receiver operating characteristic curve (AUC), calibration plots, and decision curve analysis, and SHapley Additive exPlanations (SHAP) values for interpretability. Results: The 12-month cumulative incidence of cs-irAEs was 18.0% (95% CI, 14.0%-23.0%). Nine predictors were selected, including baseline neutrophil-to-lymphocyte ratio (NLR), body mass index, prior radiotherapy, and proton pump inhibitor use. The XGBoost model achieved an AUC of 0.871 (95% CI, 0.805-0.937) and a negative predictive value (NPV) of 87.9% in the testing set. SHAP analysis revealed a non-linear protective threshold for NLR > 4.0. A bedside nomogram was created using the six strongest predictors. Conclusion: This machine learning-based model accurately identified advanced NSCLC patients at risk for cs-irAEs. Its high NPV value helped identify low-risk patients, supporting optimized monitoring and resource allocation. Combining competing risk analysis with interpretable machine learning offers a stable tool for personalized toxicity management.

Keywords: Non-small cell lung cancer, immune checkpoint inhibitors, immune-related adverse events, machine learning, prediction model

Introduction

Non-small cell lung cancer (NSCLC) is the most common cause of cancer-related deaths worldwide, with recent projections indicating approximately 2.5 million new cases and 1.8 million deaths in 2022 [1]. This landscape has been transformed by the emergence of immune checkpoint inhibitors (ICIs), which have rejuvenated the host immune response against cancer. Notable clinical trials, such as KEYNOTE-189, have demonstrated significant survival benefits, thereby establishing these agents as a cornerstone of first-line therapy in patients with advanced NSCLC without targetable driver mutations [2].

This potent immune activation is, however, a double-edged sword, associated with a broad range of off-target inflammatory toxicities, collectively termed immune-related adverse events (irAEs). ICI therapy leads to clinically significant irAEs (cs-irAEs, grade 2 or more) in roughly 40% of patients, with severe events (grade 3 or higher) occurring in almost 20% [3]. Importantly, these toxicities are managed according to con-

https://doi.org/10.62347/SHBS9371

sensus guidelines issued by organizations such as the European Society of Medical Oncology [4]; however, their development can severely impair quality of life and result in the discontinuation, delay, or compromise of subsequent treatment.

Although they have a clinical effect, accurately predicting which patients will develop severe irAEs remains an unresolved issue. Predictive biomarkers have been widely investigated, but studies on potential predictors such as PD-L1 expression, cytokine levels, and genetic information have not been consistently effective or practically useful. Therefore, according to current reviews, there is no established biomarker that can be implemented in clinical practice to stratify patients according to their risk of irAE [5]. Moreover, currently available predictive models tend to use conventional statistical approaches, which presuppose linear relationships and fail to handle high-dimensional data adequately as well as the complex and non-linear interactions among factors that characterize the immuno-oncological ecosystem.

To address this critical gap, more advanced predictive tools capable of incorporating diverse clinical and laboratory data are urgently needed. Machine learning, a branch of artificial intelligence, provides an effective approach to reveal intricate patterns in large-scale medical data that cannot be identified through traditional statistical techniques, and it has strong potential to improve prognostic assessment in oncology [6]. Consequently, our goal was not only to develop a predictive model, but also to establish a rigorous analytical framework that takes into account potentially conflicting risk factors and applies appropriate feature selection to minimize bias. Using a multidimensional dataset, we hypothesized that a machine learning model might better predict the risk of developing cs-irAEs in patients with advanced NSCLC, with the ultimate aim of optimizing monitoring strategies and improving the therapeutic index of immunotherapy.

Materials and methods

Study design and ethical statement

This single-center, retrospective cohort study was conducted to construct and validate a machine learning-based prediction model for irAEs in patients with advanced NSCLC. We consecutively enrolled adult patients diagnosed with stage IIIB-IV NSCLC who initiated treatment with PD-1/PD-L1 inhibitors (either as monotherapy or in combination therapies) at the Department of Oncology of The Second Affiliated Hospital of Nanjing Medical University from December 1, 2023, to June 30, 2025. This study was conducted in accordance with ethical standards and was approved by the Ethics Committee of The Second Affiliated Hospital of Nanjing Medical University (Approval number: 2026-KY-158-01). Given the retrospective design involving only analysis of anonymized clinical records, the requirement for informed consent was waived.

Study population

A systematic query was performed to screen all adult patients (≥ 18 years) with a diagnosis of advanced NSCLC, as indicated by the International Classification of Diseases, Tenth Revision codes. The initial list of potential participants underwent a rigorous three-stage validation process to ensure accuracy and eligibility. First, an automated algorithm filtered the cohort based on structured electronic health record data, including medication administration records and demographic fields. Second, the medical records of the shortlisted patients were independently reviewed by two trained oncologists to confirm pathologic diagnoses, staging, and treatment details. Finally, any discrepancies or cases with eligibility uncertainty were resolved by a senior chief physician through consensus.

Inclusion criteria: (1) a histologically or cytologically confirmed diagnosis of stage IIIB or IV NSCLC, staged according to the 8th edition of the American Joint Committee on Cancer Staging Manual [7]; (2) receipt of at least one dose of PD-1 or PD-L1 inhibitors, either as monotherapy or in combination with chemotherapy or anti-angiogenic agents. To ensure treatment homogeneity, all ICIs were administered according to standard approved fixed-dose or weight-based regimens (e.g., Pembrolizumab 200 mg every 3 weeks [Q3W], Nivolumab 240 mg Q2W or 480 mg Q4W, Camrelizumab 200 mg Q3W, Sintilimab 200 mg Q3W, Tislelizumab 200 mg Q3W, Atezolizumab

1200 mg Q3W, or Durvalumab 1500 mg Q4W). Patients were required to have completed at least one full administration cycle to ensure the capture of early-onset adverse events; (3) aged 18 years or older at the time of treatment initiation; (4) availability of complete baseline clinical data and longitudinal follow-up records. Exclusion criteria were applied to minimize confounding variables and included any of the following: (1) presence of other concurrent active malignancies within the past five years; (2) a known history of active autoimmune diseases or requirement for systemic corticosteroid therapy (> 10 mg/day prednisone equivalent) at baseline, as recommended by American Society of Clinical Oncology guidelines [8]; (3) a history of prior organ transplantation or allogeneic hematopoietic stem cell transplantation; or (4) a follow-up duration of less than three months, unless the primary endpoint (cs-irAEs, grade ≥ 2) or death occurred earlier.

Sample size determination

The sample size was determined based on the events per variable criterion, a widely accepted method for developing multivariable prediction models to mitigate the risk of overfitting [9]. Considering the inclusion of approximately 12 candidate predictor variables in the final multivariate model (e.g., Least Absolute Shrinkage and Selection Operator [LASSO]-selected features), a minimum of 10 outcome events is required per variable. Based on a comprehensive meta-analysis reporting the pooled incidence of irAEs in NSCLC patients treated with combination immunotherapy as approximately 40-50% for cs-irAEs (grade ≥ 2) [10], we conservatively estimated a 40% cumulative incidence rate for grade ≥ 2 irAEs in our mixed cohort. Accordingly, the required number of events was 120 (12 variables × 10 events/variable), which corresponds to a minimum total sample size of 300 patients (120 events/0.40 incidence). Therefore, our target sample size of approximately 300-350 patients was considered adequate.

Data partitioning

To facilitate model training and internal validation, the final eligible cohort was randomly partitioned into a training set (70%) and an independent testing set (30%) using a computer-generated random seed. Stratified sampling was employed to ensure a balanced distribution of the outcome events (irAEs) between the two datasets.

Data collection and outcome definition

Clinical data extraction was conducted using a dual-review process to ensure accuracy and minimize information bias. All electronic health records, including physician progress notes, nursing logs, medication administration records, and laboratory reports, were retrospectively reviewed. Data extraction was performed independently by two trained oncology fellows using a standardized case report form. Any discrepancies in data extraction were adjudicated by a senior attending oncologist through a detailed review of the original medical records. To guarantee the temporal validity of the prediction model, only baseline data available prior to the administration of the first dose of immunotherapy were collected, with laboratory values restricted to a window of 7 days pre-treatment.

The primary outcome of the study was the incidence of cs-irAEs, defined as adverse events of grade 2 or higher with a potential immunological etiology. The grading and causality assessment were strictly based on the National Cancer Institute Common Terminology Criteria for Adverse Events version 5.0 [11]. To ensure robust outcome classification, every potential irAE case underwent a consensus review by a multidisciplinary toxicity management team. All-cause mortality occurring prior to the onset of an irAE was defined as a competing risk event, given that death precludes the observation of subsequent adverse events. This distinction is crucial for the subsequent application of the Fine-Gray subdistribution hazard model.

Predictor variables

A comprehensive set of candidate predictor variables was curated based on clinical relevance and prior literature regarding immunotoxicity. Demographic and physical status variables included age, sex, body mass index (BMI), smoking history (quantified in pack-years), and Eastern Cooperative Oncology Group performance status [12]. Disease-specific characteristics comprised histological subtype (squamous vs. non-squamous), Tumor-Node-Meta-

stasis staging, number of metastatic sites (as a proxy for tumor burden), and molecular status including PD-L1 tumor proportion score (TPS) and driver mutations (e.g., Epidermal Growth Factor Receptor, Anaplastic Lymphoma Kinase). Treatment-related variables detailed the specific ICI agent, line of therapy, history of prior radiotherapy, and combination regimens (chemotherapy or anti-angiogenic agents). Notably, we also extracted data on concomitant medications used within 30 days prior to ICI initiation, specifically antibiotics, proton pump inhibitors (PPIs), statins, and baseline corticosteroids, as recent evidence suggests these agents may modulate the gut microbiome or systemic immunity, thereby influencing irAE risk [13]. Baseline laboratory biomarkers included complete blood counts and comprehensive metabolic panels. Derived inflammatory indices were calculated as follows: neutrophil-to-lymphocyte ratio (NLR), platelet-to-lymphocyte ratio, lymphocyte-to-monocyte ratio, and derived NLR [14]. Other biochemical markers included lactate dehydrogenase, albumin, C-reactive protein, thyroid-stimulating hormone (TSH), and tumor markers (carcinoembryonic antigen, cytokeratin 19 fragment).

Model development, validation, and interpretation

Data partitioning and preprocessing: To ensure an unbiased evaluation of model generalizability, the final eligible cohort was randomly partitioned into a training set (70%) and an independent testing set (30%) using a computer-generated random seed. Stratified sampling was employed to preserve the distribution of the outcome (cs-irAEs) across both datasets. Rigorous data preprocessing was implemented prior to model training. Variables with a missing rate exceeding 20% were excluded to prevent noise introduction. For variables with less than 20% missingness, we assumed data were missing at random and utilized the Multiple Imputation by Chained Equations algorithm to generate five complete imputed datasets [15]. The specific missing rates for these variables - primarily including TSH (15.6%), C-Reactive Protein (12.5%), Albumin (5.4%), and BMI (3.2%) - are thoroughly documented in the Table S1. To validate the robustness of our imputation strategy, a consistency test was performed. As detailed in Table S1, no statistically significant differences were observed in the central tendencies and dispersions of these variables before and after the imputation process (all $P > 0.05$), indicating high imputation quality without introducing artificial distributional shifts. Continuous variables were standardized using Z-score normalization (mean = 0, standard deviation = 1) to ensure feature parity and facilitate the convergence of gradient-based algorithms. Crucially, to mitigate the bias arising from the imbalanced class distribution of irAEs, the Synthetic Minority Over-sampling Technique (SMOTE) was applied exclusively to the training set [16]. Specifically, we configured the SMOTE algorithm with the k-neighbors parameter set to 5 and an oversampling ratio designed to achieve a 1:1 balanced distribution between the minority (cs-irAEs) and majority classes. To stringently control for the potential introduction of noise data and the subsequent risk of overfitting, the synthetic generation was strictly confined to the training phase. The generalization capability of the models was continuously monitored using a 5-fold cross-validation strategy during hyperparameter tuning, and ultimately validated on the independent, unmodified testing set, which preserved the real-world incidence rate and data integrity. This technique synthesized minority class examples by interpolating between existing positive samples, ensuring that the model learned robust decision boundaries without compromising the integrity of the unseen testing data.

The five imputed datasets were handled separately to preserve the uncertainty estimation of missing data. For each imputed dataset, SMOTE was applied to its training subset. Subsequently, feature selection (LASSO) and model training were performed independently on each of the five augmented training sets. For prediction on the testing set, the final risk score for each patient was computed as the average of the predicted probabilities from the five models corresponding to the five imputed datasets.

Holistic feature selection: To construct a parsimonious predictive signature while strictly avoiding the “double selection bias” associated with pre-screening based on univariate *P*-values, we adopted a holistic feature selection strategy. All collected clinically relevant variables spanning demographics, disease characteristics, treatment details, and blood biomark-

ers were directly input into a LASSO logistic regression model. The optimal regularization parameter (λ) was determined by stratified 10-fold cross-validation on the training set, utilizing binomial deviance as the specific loss function to account for the binary nature of our clinical outcome. Variables retaining non-zero coefficients at the optimal λ (using the 1-standard-error rule to favor model simplicity) were identified as the final predictor set for subsequent model development [17].

Model construction and optimization: Based on the LASSO-selected features, four distinct supervised machine learning algorithms were developed: Logistic Regression as a linear baseline, and three non-linear models including Random Forest, Extreme Gradient Boosting (XGBoost), and Support Vector Machine. Hyperparameter tuning was conducted to optimize model performance using a comprehensive grid search method combined with 5-fold cross-validation on the training set. Specifically, the tuning objective function was set to maximize the cross-validated Area Under the Receiver Operating Characteristic Curve (AUC), utilizing "binary:logistic" as the learning task parameter for the ensemble models. The predefined parameter search space for the XGBoost model included learning_rate (0.01, 0.05, 0.1, 0.2), max_depth (3, 4, 5, 6, 7), and n_estimators (50 to 300 with a step of 50). Similarly, for the Random Forest model, the grid search encompassed n_estimators (100 to 500 with a step of 100) and max_depth (5 to 15). The detailed optimization trajectories and corresponding validation curves, which illustrate the relationship between key hyperparameters and model performance (AUC), have been supplemented in Figure S1. Key hyperparameters, such as the number of estimators and maximum depth for tree-based models (Random Forest and XGBoost) and the penalty parameter (C) for Support Vector Machine, were iteratively adjusted to maximize the AUC.

Model performance evaluation: Comprehensive performance evaluation was conducted on the independent testing set. Discrimination was assessed using the AUC, with the DeLong test employed to compare the statistical significance of AUC differences between models. Additional metrics, including accuracy, sensitivity, specificity, positive predictive value, negative predictive value, F1-score, as well as the clinically actionable positive likelihood ratio (+LR) and negative likelihood ratio (-LR), were calculated at the probability threshold that maximized the Youden index in the training set. The likelihood ratios were incorporated to directly translate the model's discriminative ability into evidence-based clinical decision-making metrics. Calibration was evaluated by visual inspection of calibration plots and the Hosmer-Lemeshow goodness-of-fit test ($P > 0.05$ indicating good agreement between predicted and observed probabilities). Furthermore, the clinical utility of the optimal model was quantified using Decision Curve Analysis (DCA), which estimates the net benefit across a range of clinically relevant risk thresholds. In addition, to assess the robustness of the optimal XGBoost model, a subgroup analysis was conducted across key clinical strata, including age (> 65 vs. ≤ 65 years), PD-L1 expression status (< 1% vs. ≥ 1%), specific ICI agent types (PD-1 vs. PD-L1 inhibitors), and treatment regimens (monotherapy vs. combination therapy).

Interpretation and visualization of models: In order to explain the decision-making mechanism of the "black-box" ensemble models, we used SHapley Additive exPlanations (SHAP) [18]. Specifically, for our optimal XGBoost model, the exact SHAP values were computed utilizing the computationally efficient TreeExplainer algorithm, which is purposefully designed for tree-based machine learning models. The statistical basis for ranking global feature importance was determined by calculating the mean absolute SHAP value for each predictor across all patients in the dataset, representing the average magnitude of a feature's impact on the model's predictive output. SHAP summary plots were created to depict the overall significance and directional effects of each feature on the predicted risk of irAEs. A hierarchical visualization strategy was employed in order to overcome the trade-off between predictive accuracy and clinical usability. If the multivariate logistic regression model showed competitive performance, i.e., an AUC within 0.02 of the best-performing non-linear model, its coefficients were used to build a fixed nomogram because it offers transparency and is easy to use. On the other hand, when the ensemble models, e.g., XGBoost performed much better than the logistic regression baseline, SHAP

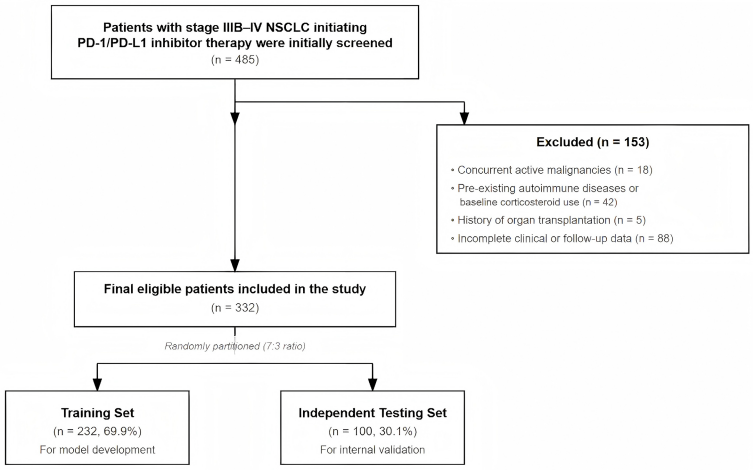

Figure 1. Flow diagram. Abbreviations: NSCLC, Non-Small Cell Lung Cancer; PD-1, Programmed Cell Death 1; PD-L1, Programmed Death-Ligand 1; n, number of patients.

dependence plots and SHAP summary plots were used as the leading clinical interpretation tools to ensure that simplicity was not compromised for predictive precision.

Statistical analysis

Data processing was performed using R software (version 4.5.0, Vienna, Austria) and Python (version 3.9). We first evaluated the normality of continuous variables by the Shapiro-Wilk test. Accordingly, data were reported as mean ± standard deviation for normal distributions or median with interquartile range (IQR) for non-normal distributions. Categorical data were described as counts and percentages [n (%)]. To assess baseline balance between the training and independent testing cohorts, we applied the independent samples t-test or Mann-Whitney U test for continuous data, and the Chi-square test or Fisher's exact test for categorical data. Given the potential competing risk of death, the Fine-Gray subdistribution hazard model was utilized to analyze outcome-associated factors. The observation period started from the first infusion of ICIs and continued until the first documented occurrence of a cs-irAE (Grade ≥ 2), all-cause death, or the last date of clinical follow-up (administrative censoring), whichever occurred first. All-cause death occurring before the onset of cs-irAE (grade ≥ 2) was defined as a competing event. To verify the proportional subdistribution hazards assumption of the Fine-Gray model, we evaluated the Schoenfeld-type residuals, and the global test indicated no significant violation ($P > 0.05$). Furthermore, to explicitly demonstrate the effect of the competing risk of death, a standard cause-specific Cox proportional hazards model - which treats death as independent, non-informative censoring - was performed as a sensitivity analysis. We calculated Subdistribution Hazard Ratios (SHR) and 95% Confidence Intervals (CI) through univariate Fine-Gray analysis to estimate the raw association between baseline variables and the risk of cs-irAEs, and subsequently compared these estimates with the Hazard Ratios derived from the standard Cox model. Notably, these univariate results were used for clinical interpretation only and not for variable selection, ensuring no conflict with the subsequent LASSO modeling. All tests were two-sided, with significance set at $P < 0.05$.

Results

Patient screening flow and baseline characteristics

As illustrated in the study flowchart (**Figure 1**), a total of 485 patients with stage IIIB-IV NSCLC initiating PD-1/PD-L1 inhibitor therapy were initially screened. After excluding patients with concurrent active malignancies (n = 18), pre-existing autoimmune diseases or baseline corticosteroid use (n = 42), history of organ transplantation (n = 5), and incomplete clinical or follow-up data (n = 88), a final cohort of 332 eligible patients was included in the analysis. The median follow-up duration for the entire cohort was 14.5 months (IQR, 8.2-22.4 months). Using a computer-generated random seed, the cohort was partitioned into a training set (n = 232, 69.9%) for model development and an independent testing set (n = 100, 30.1%) for internal validation. The baseline demographic, clinicopathological, treatment-related, and laboratory characteristics of patients in both sets are summarized in **Table 1**. The mean age of the total population was 63.4 ± 9.8 years, with a predominance of male patients (65.4%) and non-squamous histology

Table 1. Baseline characteristics of patients in the training and independent testing sets

Characteristic	Training Set (n = 232)	Independent Testing Set (n = 100)	Statistic ($t/Z/\chi^2$)	*P*-value (Between-group)	Shapiro-Wilk Test (W, *P*-value)*
Demographics					
Age (years), Mean ± SD	63.1 ± 9.5	64.0 ± 10.2	t = -0.785	0.433	W = 0.985, P = 0.214
Sex (Male/Female), n (%)	150 (64.7)/82 (35.3)	67 (67.0)/33 (33.0)	χ^2 = 0.184	0.668	-
BMI (kg/m^2), Mean ± SD	23.4 ± 3.2	23.1 ± 3.5	t = 0.742	0.459	W = 0.978, P = 0.152
Smoking history (pack-years), Median [IQR]	25 [0, 40]	22.5 [0, 45]	Z = 0.356	0.722	W = 0.852, P < 0.001
ECOG PS (0/1/≥ 2), n (%)	78 (33.6)/135 (58.2)/19 (8.2)	36 (36.0)/56 (56.0)/8 (8.0)	χ^2 = 0.542	0.763	-
Disease Characteristics					
Histology (Non-squamous/Squamous), n (%)	166 (71.6)/66 (28.4)	70 (70.0)/30 (30.0)	χ^2 = 0.095	0.758	-
TNM Stage (IIIB-C/IV), n (%)	54 (23.3)/178 (76.7)	21 (21.0)/79 (79.0)	χ^2 = 0.257	0.612	-
Metastatic sites (n), Median [IQR]	2 [1, 3]	2 [1, 3]	Z = -0.124	0.901	W = 0.784, P < 0.001
Driver mutation (Wild-type/Positive), n (%)	184 (79.3)/48 (20.7)	81 (81.0)/19 (19.0)	χ^2 = 0.312	0.856	-
PD-L1 TPS (< 1%/1-49%/≥ 50%), n	85/94/53	34/43/23	χ^2 = 0.455	0.797	-
Treatment Characteristics					
ICI Agent Type (PD-1/PD-L1), n (%)	182 (78.4)/50 (21.6)	80 (80.0)/20 (20.0)	χ^2 = 0.118	0.731	-
Specific ICI Agent^, n (%)			χ^2 = 1.052	0.902	-
Pembrolizumab	75 (32.3)	33 (33.0)			-
Camrelizumab/Sintilimab/Tislelizumab	107 (46.1)	47 (47.0)			-
Atezolizumab/Durvalumab	50 (21.6)	20 (20.0)			-
Treatment Regimen (Mono/Combo), n	85/147	39/61	χ^2 = 0.166	0.684	-
Line of Therapy (1st/≥ 2nd), n	158/74	69/31	χ^2 = 0.065	0.799	-
Prior Radiotherapy (Yes), n (%)	72 (31.0)	28 (28.0)	χ^2 = 0.298	0.585	-
Concomitant Medications (within 30 days)					
Antibiotics use, n (%)	45 (19.4)	18 (18.0)	χ^2 = 0.091	0.763	-
Proton Pump Inhibitors (PPIs) use, n (%)	68 (29.3)	32 (32.0)	χ^2 = 0.237	0.626	-
Statins use, n (%)	38 (16.4)	15 (15.0)	χ^2 = 0.104	0.747	-
Baseline Corticosteroids (< 10 mg), n (%)	12 (5.2)	5 (5.0)	χ^2 = 0.006	0.939	-
Tumor Markers					
CEA (ng/mL), Median [IQR]	15.2 [5.1, 45.3]	14.8 [4.8, 42.1]	Z = -0.112	0.911	W = 0.655, P < 0.001
CYFRA21-1 (ng/mL), Median [IQR]	4.2 [2.1, 8.5]	4.5 [2.3, 8.8]	Z = -0.455	0.649	W = 0.712, P < 0.001
Baseline Laboratory Biomarkers					
WBC count (× 10^9/L), Mean ± SD	6.8 ± 2.1	6.9 ± 2.3	t = -0.384	0.701	W = 0.962, P = 0.085
Hemoglobin (g/L), Mean ± SD	115.4 ± 16.2	114.1 ± 15.8	t = 0.687	0.493	W = 0.975, P = 0.128
Albumin (g/L), Mean ± SD	38.5 ± 4.2	39.0 ± 4.5	t = -0.965	0.335	W = 0.981, P = 0.334
LDH (U/L), Median [IQR]	215 [178, 280]	220 [182, 275]	Z = -0.412	0.680	W = 0.825, P < 0.001
CRP (mg/L), Median [IQR]	8.5 [3.2, 18.4]	9.1 [3.5, 19.2]	Z = -0.558	0.577	W = 0.758, P < 0.001
TSH (mIU/L), Median [IQR]	2.1 [1.4, 3.2]	2.2 [1.3, 3.1]	Z = -0.215	0.830	W = 0.884, P < 0.001
Inflammatory Indices					
NLR, Median [IQR]	3.2 [2.1, 4.8]	3.4 [2.2, 5.0]	Z = -0.842	0.399	W = 0.865, P < 0.001
PLR, Median [IQR]	165 [120, 230]	170 [125, 240]	Z = -0.635	0.525	W = 0.892, P < 0.001
LMR, Median [IQR]	3.5 [2.4, 4.9]	3.3 [2.5, 4.8]	Z = 0.488	0.625	W = 0.905, P < 0.001
dNLR, Median [IQR]	2.1 [1.5, 3.0]	2.2 [1.6, 3.1]	Z = -0.512	0.609	W = 0.873, P < 0.001

Notes: *The Shapiro-Wilk test for normality was performed on the total cohort (N = 332) for all continuous variables. A *P*-value < 0.05 indicates a significant deviation from a normal distribution, thereby statistically justifying the use of Median [IQR] and the non-parametric Mann-Whitney U test (Z statistic). ^Specific ICI agents were supplemented to ensure transparency of treatment regimens. Abbreviations: BMI, Body Mass Index; CEA, Carcinoembryonic Antigen; CRP, C-Reactive Protein; CYFRA21-1, Cytokeratin-19 Fragment; dNLR, derived Neutrophil-to-Lymphocyte Ratio; ECOG PS, Eastern Cooperative Oncology Group Performance Status; ICI, Immune Checkpoint Inhibitor; IQR, Interquartile Range; LDH, Lactate Dehydrogenase; LMR, Lymphocyte-to-Monocyte Ratio; NLR, Neutrophil-to-Lymphocyte Ratio; PD-1, Programmed Cell Death 1; PD-L1, Programmed Death-Ligand 1; PLR, Platelet-to-Lymphocyte Ratio; PPIs, Proton Pump Inhibitors; SD, Standard Deviation; TNM, Tumor, Node, Metastasis; TPS, Tumor Proportion Score; TSH, Thyroid-Stimulating Hormone; WBC, White Blood Cell.

(71.1%). The majority of patients (62.7%) received immune checkpoint inhibitors in combination regimens, and most were treated in the first-line setting. The detailed distribution of specific ICI agents has been newly incorporated into the revised baseline characteristics. Fur-

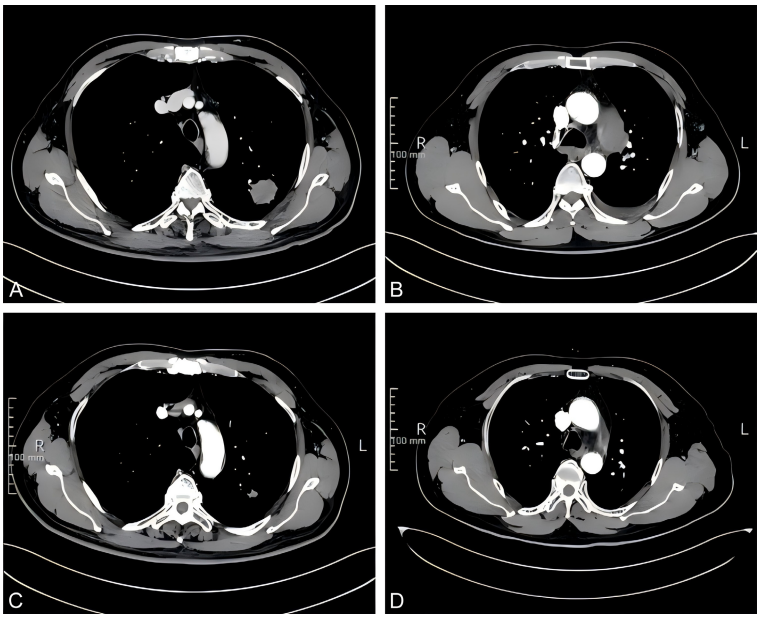

Figure 2. Representative CT scans demonstrating radiological response to PD-1/PD-L1 inhibitor therapy. (A, B) Baseline chest CT images obtained prior to immunotherapy, revealing a primary soft-tissue mass in the left upper lobe (A) and significant metastatic mediastinal lymphadenopathy (B). (C, D) Follow-up CT scans after treatment initiation, showing marked regression of both the primary pulmonary lesion (C) and the mediastinal lymph nodes (D), indicative of a partial response. Abbreviations: CT, Computed Tomography; PD-1, Programmed Cell Death 1; PD-L1, Programmed Death-Ligand 1.

thermore, to strictly justify our data presentation formats (mean ± standard deviation vs. median with IQR), the Shapiro-Wilk normality test was performed for all continuous variables across the total cohort. The specific test statistics (W) and corresponding *P*-values confirmed the non-normal distribution of several key variables, such as smoking history and inflammatory indices (all Shapiro-Wilk $P < 0.05$), which are explicitly detailed in **Table 1**. Statistical comparison revealed no significant differences between the training and independent testing sets across all analyzed variables (all $P > 0.05$), confirming that the randomization process successfully achieved a balanced distribution of covariates and minimized selection bias.

Therapeutic efficacy and representative clinical case

Although the primary focus of this study was to model the risk of adverse events, it is clinically relevant to acknowledge the therapeutic efficacy observed in our cohort. Consistent with the established benefits of ICIs, the majority of patients exhibited favorable radiological responses. A representative case is illustrated in **Figure 2**, demonstrating marked regression of both the primary pulmonary mass and mediastinal lymphadenopathy following the initiation of PD-1/PD-L1 inhibitor therapy. This context highlights the potent immune activation that underlies both the therapeutic benefit and the potential for toxicity.

Epidemiology, risk factors, and clinical outcomes of irAEs

In the total cohort of 332 patients, all-cause mortality occurred in 58 patients prior to the documented onset of any irAEs, underscoring the necessity of accounting for death as a competing risk. Based on the Fine-Gray subdistribution hazard model, the cumulative incidence of cs-irAEs (Grade ≥ 2) exhibited a progressive increase over the treatment course. As illustrated in **Figure 3A**, the estimated cumulative incidence rates of cs-irAEs at 3, 6, and 12 months were 3.8% (95% CI, 2.1%-6.3%), 9.6% (95% CI, 6.6%-13.0%), and 18.0% (95% CI, 14.0%-23.0%), respectively. The median time to the onset of the first cs-irAE was 5.2 months (IQR, 2.5-9.8 months), with the majority of events occurring within the first year of therapy.

The spectrum and severity distribution of the identified irAEs are visualized in **Figure 3B**. Among the patients who developed irAEs (n = 93), skin toxicities including rash and pruritus were the most prevalent, accounting for 35.5% (33 cases) of all recorded events, followed by endocrine disorders such as hypothyroidism and thyroiditis at 23.7% (22 cases). The majority of these dermatological and endocrine events were classified as Grade 2. In contrast, high-grade toxicities (Grade 3-5) were predominantly observed in patients diagnosed with checkpoint inhibitor pneumonitis (12.9% of events) and immune-mediated hepatitis (8.6% of events), highlighting the organ-specific het-

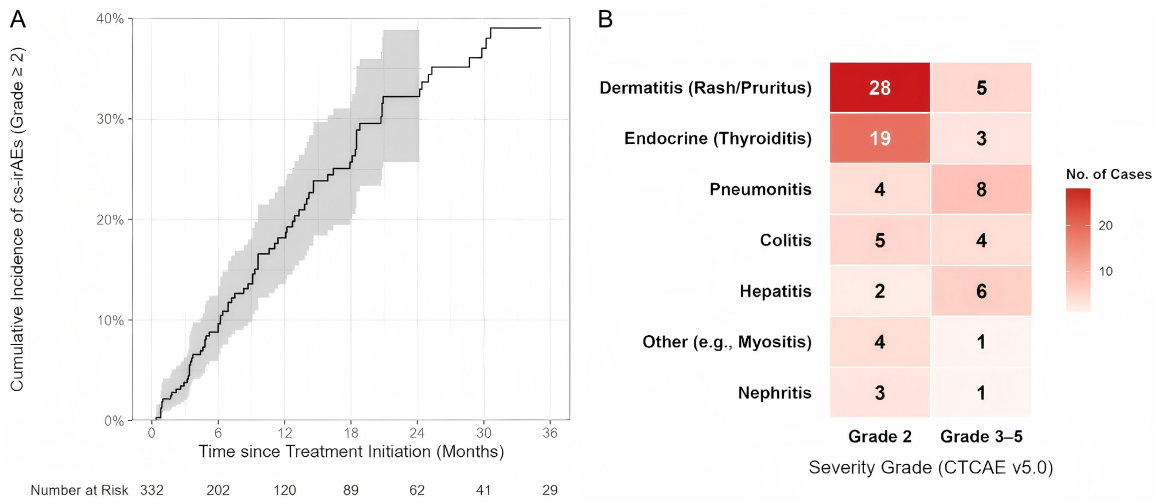

Figure 3. Temporal incidence and clinical spectrum of cs-irAEs. A. CIF curve illustrating the probability of developing cs-irAEs (Grade ≥ 2) over time in the overall cohort (N = 332), accounting for all-cause mortality as a competing risk event. The shaded area represents the 95% confidence interval. The table below the graph indicates the number of patients at risk at specific time points following treatment initiation. B. Heatmap visualization displaying the frequency and severity distribution of identified irAEs (n = 93 total events) across different organ systems. Rows represent specific organ toxicities sorted by total frequency from bottom to top, and columns represent the severity grade according to the CTCAE v5.0 (Grade 2 vs. Grade 3-5). Color intensity and numerical labels within each cell indicate the absolute number of cases in that category. Abbreviations: cs-irAEs, Clinically Significant Immune-Related Adverse Events; CTCAE, Common Terminology Criteria for Adverse Events; CIF, Cumulative Incidence Function.

erogeneity in toxicity severity. Regarding the clinical management and prognosis of these recorded events, all therapeutic interventions strictly adhered to established consensus guidelines (e.g., American Society of Clinical Oncology/European Society of Medical Oncology). Specifically, the majority of Grade 2 events (such as dermatitis and thyroiditis) were successfully managed with temporary ICI interruption, low-dose oral corticosteroids, or targeted hormone replacement. Conversely, severe Grade 3-5 toxicities necessitated prompt ICI discontinuation and high-dose intravenous systemic corticosteroid therapy, with secondary immunosuppressants (e.g., mycophenolate mofetil) required in a small subset of steroid-refractory cases. Overall, the clinical outcomes were favorable, with approximately 85.0% of the observed cs-irAEs resolving or improving to Grade ≤ 1 following appropriate clinical intervention, underscoring the reversibility of most events when detected and managed in a timely manner. To further elucidate the clinical trajectory of these toxicities, the median onset time and specific resolution outcomes stratified by each distinct organ system have been detailed in Table S3.

To explore the crude associations between baseline characteristics and the risk of cs-irAEs, a univariate Fine-Gray analysis was performed in the total cohort (**Table 2**). Several clinical factors demonstrated statistically significant associations with an altered risk of irAEs. Specifically, a history of prior radiotherapy (SHR = 1.62; 95% CI, 1.05-2.48; P = 0.029), the use of combination therapy regimens compared to monotherapy (SHR = 1.55; 95% CI, 1.02-2.36; P = 0.041), and a higher baseline BMI (SHR = 1.08 per unit increase; 95% CI, 1.02-1.14; P = 0.009) were identified as risk factors. Conversely, a higher baseline NLR was significantly associated with a reduced risk of irAEs (SHR = 0.85; 95% CI, 0.76-0.95; P = 0.005), suggesting an inverse relationship between baseline systemic inflammation and the development of immunotoxicity. Notably, a comparison between the Fine-Gray model and the standard Cox proportional hazards model revealed that ignoring the competing risk of death systematically overestimated the hazard estimates for most significant predictors (detailed in Table S2), thereby confirming the analytical necessity of our competing risk adjustment strategy.

Table 2. Univariate Fine-Gray analysis of baseline factors associated with cs-irAEs in the total cohort (N = 332)

Variable	Events in Category, n	SHR	95% CI	*P*-value
Demographics				
Age (per 5-year increase)	-	1.04	0.95-1.14	0.382
Sex (Male vs. Female)	62	0.89	0.58-1.36	0.591
BMI (per 1 kg/m^2 increase)	-	1.08	1.02-1.14	0.009
ECOG PS (1 vs. 0)	48	1.12	0.74-1.69	0.588
Disease Characteristics				
Histology (Squamous vs. Non-squamous)	30	1.21	0.78-1.88	0.395
TNM Stage (IV vs. IIIB/C)	69	1.15	0.72-1.84	0.554
PD-L1 TPS (≥ 1% vs. < 1%)	55	1.34	0.88-2.04	0.172
Treatment Characteristics				
Therapy Type (Combination vs. Mono)	66	1.55	1.02-2.36	0.041
Prior Radiotherapy (Yes vs. No)	36	1.62	1.05-2.48	0.029
Concomitant Medications				
Antibiotics use (Yes vs. No)	13	0.75	0.42-1.35	0.341
Proton Pump Inhibitors (Yes vs. No)	27	1.05	0.65-1.68	0.845
Biomarkers				
Albumin (per 5 g/L decrease)	-	1.18	0.94-1.48	0.154
NLR (per 1-unit increase)	-	0.85	0.76-0.95	0.005
dNLR (per 1-unit increase)	-	0.82	0.71-0.94	0.006

Notes: Univariate Fine-Gray competing risk regression analysis was performed to identify baseline factors associated with the cumulative incidence of cs-irAEs. Death without prior cs-irAEs was treated as a competing event. For continuous variables (e.g., Age, BMI, Albumin, NLR, dNLR), the number of events in specific categories is not applicable (-) as the SHR represents the hazard change per unit increase/decrease. Abbreviations: BMI, Body Mass Index; CI, Confidence Interval; cs-irAEs, Clinically Significant Immune-Related Adverse Events; dNLR, derived Neutrophil-to-Lymphocyte Ratio; ECOG PS, Eastern Cooperative Oncology Group Performance Status; NLR, Neutrophil-to-Lymphocyte Ratio; PD-L1, Programmed Death-Ligand 1; SHR, Subdistribution Hazard Ratio; TNM, Tumor, Node, Metastasis; TPS, Tumor Proportion Score.

Holistic feature selection and multivariate analysis

To identify a parsimonious set of robust predictors for model development while strictly avoiding double selection bias, feature selection was performed exclusively on the training set (n = 232) using a holistic strategy. Instead of prefiltering variables based on univariate *P*-values, all 26 collected clinically relevant variables were directly entered into a LASSO logistic regression. The coefficient profiles of the variables across a range of penalty terms (log(λ)) are shown in **Figure 4A**. The optimal λ value (λ = 0.034) was determined by 10-fold cross-validation, corresponding to the minimum binomial deviance (**Figure 4B**). At this optimal threshold (λ = 0.034), nine predictors retained non-zero coefficients and were selected for inclusion in the subsequent model-building phase. The specific penalized coefficients for these selected features were as follows: history of prior radiotherapy (0.615), combination therapy regimen (0.482), concomitant use of PPIs (0.324), baseline BMI (0.058), TSH (0.081), PD-L1 TPS ≥ 1% (0.215), histological subtype of squamous cell carcinoma (0.156), antibiotic exposure (0.112), and NLR (-0.185).

To obtain clinically interpretable effect estimates for these selected features, the nine variables were entered into a traditional multivariable logistic regression model within the training set. The results of this final multivariate analysis are presented in **Table 3**. Consistent with the LASSO selection, key variables were confirmed as independent predictors of cs-irAEs. A history of prior radiotherapy (aOR = 2.15; 95% CI: 1.24-3.75; P = 0.007) and the use of combination therapy (aOR = 1.88; 95% CI: 1.12-3.18; P = 0.018) were strong risk factors. Notably, concomitant use of PPIs emerged as a significant predictor (aOR = 1.65; 95% CI: 1.05-2.60; P = 0.031). Additionally, higher

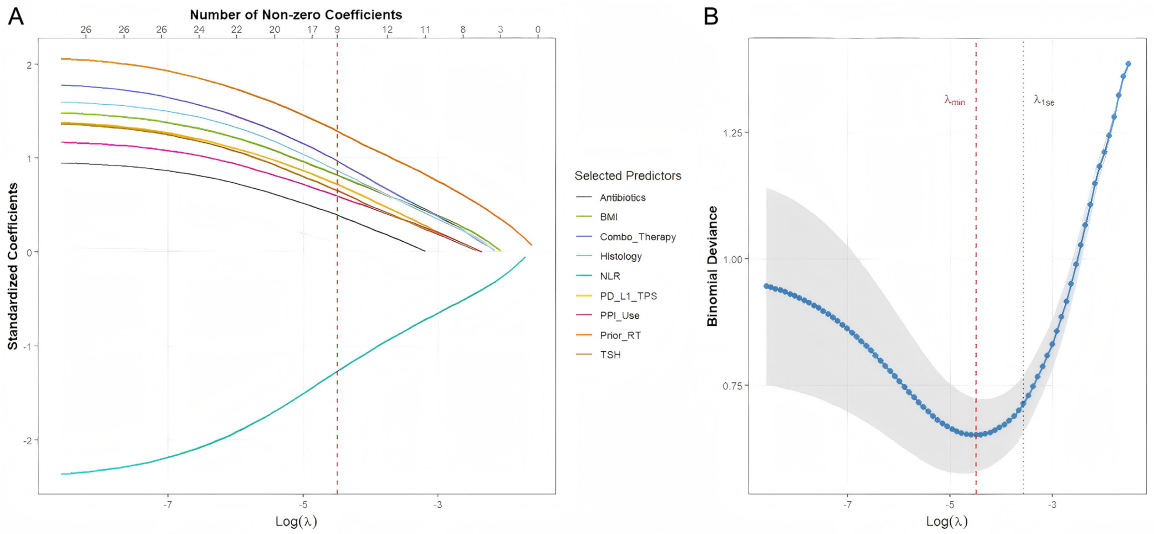

Figure 4. Feature selection via LASSO regression. A. LASSO coefficient profiles of the 26 candidate clinical variables. Each colored line represents the coefficient trajectory of a specific variable as the tuning parameter λ (log scale) changes. B. Selection of the optimal penalization coefficient λ using 10-fold cross-validation. Abbreviations: BMI, Body Mass Index; LASSO, Least Absolute Shrinkage and Selection Operator; NLR, Neutrophil-to-Lymphocyte Ratio; PD-L1, Programmed Death-Ligand 1; PPI, Proton Pump Inhibitor; RT, Radiotherapy; TPS, Tumor Proportion Score; TSH, Thyroid-Stimulating Hormone.

Table 3. Multivariate logistic regression analysis of independent predictors for cs-irAEs identified by LASSO in the training set

Predictor	β	SE	Wald χ^2	aOR	95% CI for aOR	*P*-value
Demographics						
BMI (per 1 kg/m^2 increase)	0.095	0.042	5.12	1.10	1.01-1.19	0.024
Disease Characteristics						
PD-L1 TPS (≥ 1% vs. < 1%)	0.451	0.285	2.5	1.57	0.90-2.75	0.113
Histology (Squamous vs. Non-squamous)	0.320	0.310	1.06	1.38	0.75-2.53	0.302
Treatment Characteristics						
Prior Radiotherapy (Yes vs. No)	0.765	0.281	7.41	2.15	1.24-3.75	0.007
Therapy Type (Combo vs. Mono)	0.631	0.266	5.62	1.88	1.12-3.18	0.018
Concomitant Medications						
PPI Use (Yes vs. No)	0.501	0.232	4.66	1.65	1.05-2.60	0.031
Antibiotics Use (Yes vs. No)	0.412	0.298	1.91	1.51	0.84-2.70	0.166
Biomarkers						
NLR (per 1-unit increase)	-0.248	0.069	12.92	0.78	0.68-0.89	< 0.001
TSH (per 1 mIU/L deviation)	0.115	0.052	4.89	1.12	1.01-1.24	0.027

Notes: The model was constructed using the training set (n = 232) and includes all nine variables selected by LASSO regression. The overall model fit was $\chi^2 = 45.2$, $P < 0.001$, Nagelkerke $R^2 = 0.38$. Abbreviations: aOR, Adjusted Odds Ratio; BMI, Body Mass Index; CI, Confidence Interval; cs-irAEs, Clinically Significant Immune-Related Adverse Events; LASSO, Least Absolute Shrinkage and Selection Operator; NLR, Neutrophil-to-Lymphocyte Ratio; PD-L1, Programmed Death-Ligand 1; PPI, Proton Pump Inhibitor; SE, Standard Error; TPS, Tumor Proportion Score; TSH, Thyroid-Stimulating Hormone; β, Regression Coefficient.

baseline BMI (aOR = 1.10 per 1 kg/m^2 increase; 95% CI: 1.01-1.19; P = 0.024) and deviations in TSH levels (aOR = 1.12 per 1 mIU/L deviation; 95% CI: 1.01-1.24; P = 0.027) independently augmented the risk of developing cs-irAEs. Conversely, higher baseline NLR remained an independent protective factor (aOR = 0.78 per 1-unit increase; 95% CI: 0.68-0.89; P < 0.001).

Table 4. Hyperparameter optimization and model development characteristics on the training set

Algorithm	Optimal Hyperparameters	Mean CV-AUC (95% CI)	Top 3 Predictors (Importance Metric)	Model Complexity/ Notes
LR	Penalty: L2 (Ridge); C: 0.1; Solver: liblinear	0.781 (0.735-0.827)	Metric: Coefficient Magnitude; 1. Prior Radiotherapy; 2. Combination Therapy; 3. PPI Use	10 coefficients (incl. intercept)
RF	n_estimators: 200; max_depth: 8; min_samples_split: 5	0.845 (0.802-0.888)	Metric: Gini Importance; 1. NLR; 2. BMI; 3. Prior Radiotherapy	~30,000 nodes total
XGBoost	learning_rate: 0.05; max_depth: 4; subsample: 0.8; n_estimators: 150	0.862 (0.825-0.899)	Metric: Gain; 1. NLR; 2. Prior Radiotherapy; 3. TSH	~1,200 trees total

Notes: CV-AUC: Mean area under the ROC curve with 95% confidence interval from the 10-fold cross-validation performed during grid search on the training set (n = 232). All models were fitted using the same set of nine predictors selected by LASSO. Abbreviations: |Coefficient|, absolute value of regression coefficient; Gini Importance, mean decrease in impurity; Gain, average gain across all splits the feature is used in. Abbreviations: AUC, Area Under the Curve; C, Inverse of Regularization Strength; CI, Confidence Interval; CV, Cross-Validation; LASSO, Least Absolute Shrinkage and Selection Operator; LR, Logistic Regression; NLR, Neutrophil-to-Lymphocyte Ratio; PPI, Proton Pump Inhibitor; RF, Random Forest; ROC, Receiver Operating Characteristic; TSH, Thyroid-Stimulating Hormone; XGBoost, Extreme Gradient Boosting.

Optimal model configurations and development performance

The three machine learning models were trained and optimized on the training set using the nine predictors selected by LASSO. **Table 4** summarizes the hyperparameter optimization process and key development characteristics for each algorithm. The XGBoost model achieved the highest mean cross-validated AUC (CV-AUC = 0.862, 95% CI: 0.825-0.899) during development, demonstrating robustness in handling complex feature interactions. The top three predictors for each model, identified based on algorithm-specific importance metrics, are presented within **Table 4**. Key clinical variables including prior radiotherapy, NLR, and BMI emerged as top contributors, reinforcing their pivotal role in predicting immunotoxicity across different modeling approaches.

Multidimensional performance validation in independent testing set

The generalization capability of the developed models was rigorously evaluated in the independent testing set (n = 100). As illustrated in **Figure 5**, the XGBoost model demonstrated the highest discrimination performance, yielding an AUC of 0.871 (95% CI: 0.805-0.937). This was followed by the Random Forest model with an AUC of 0.849 (95% CI: 0.778-0.920). The traditional Logistic Regression model, while effective, showed a comparatively lower AUC of 0.795 (95% CI: 0.712-0.878). Statistical comparison using the DeLong test confirmed that the XGBoost model significantly outperformed the Logistic Regression baseline (Z = 2.074, P = 0.038), justifying the use of advanced machine learning algorithms to capture non-linear interactions among clinical features.

Comprehensive performance metrics at the optimal probability thresholds are detailed in **Table 5**. The XGBoost model achieved the best overall balance between sensitivity (81.6%) and specificity (82.3%), resulting in a high F1-score of 0.782. Notably, the model exhibited a strong negative predictive value of 87.9%. Furthermore, to enhance its utility in clinical decision-making, the XGBoost model demonstrated a robust positive likelihood ratio (+LR = 4.61) and a substantially low negative likelihood ratio (-LR = 0.22). An -LR value of 0.22 effectively generates a meaningful decrease in the pre-test probability of immunotoxicity, further indicating its reliability in correctly identifying patients at low risk of developing cs-irAEs and optimizing healthcare resource allocation.

To assess further the robustness of the optimal XGBoost model, a subgroup analysis was conducted across key clinical strata (**Table 6**). The predictive performance remained stable across disparate demographic and disease subgroups. Specifically, the model maintained robust discrimination in elderly patients (> 65 years: AUC = 0.854), patients with squamous histology

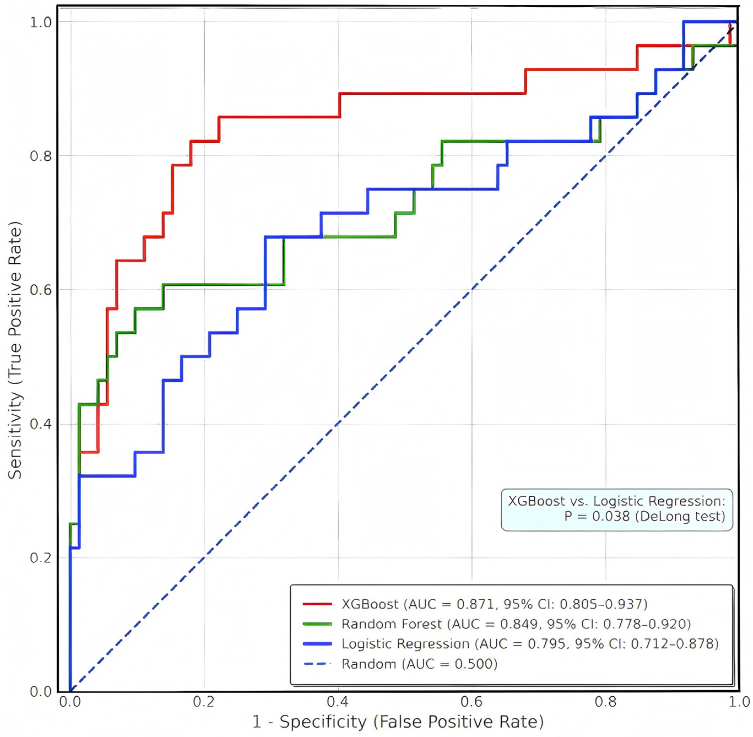

Figure 5. Comparison of ROC curves in independent testing set. ROC curves illustrates the discrimination performance of the XGBoost (red), Random Forest (green), and Logistic Regression (blue) models in the independent testing set (n = 100). The XGBoost model demonstrated a significantly higher AUC compared to the Logistic Regression model (DeLong test P = 0.038). Abbreviations: ROC, Receiver operating characteristic; AUC, Area Under the Curve; CI, confidence interval; XGBoost, Extreme Gradient Boosting.

(AUC = 0.865), and those with PD-L1 TPS < 1% (AUC = 0.880). Crucially, the XGBoost model also demonstrated excellent stability across different therapeutic modalities, achieving an AUC of 0.869 (95% CI: 0.782-0.956) in patients receiving PD-1 inhibitors and 0.862 (95% CI: 0.771-0.953) in those treated with combination therapy regimens, thereby confirming its broad applicability in diverse clinical scenarios.

Model interpretation and clinical use

To bridge the gap between algorithmic predictions and clinical decision-making, we conducted a thorough analysis of the calibration and clinical net benefit of the optimal XGBoost model. The calibration curve of the XGBoost model in **Figure 6A** indicated that the model worked well in predicting the probabilities and the observed frequencies of cs-irAEs in the independent testing set. The *P*-value of the Hosmer-Lemeshow test (P = 0.285) was non-significant, which is sufficient evidence of adequate calibration of the best-performing model. Moreover, DCA found that applying the XGBoost model to inform the selection of monitoring strategies would have a higher net benefit than either the treat-all or treat-none strategies over a broad spectrum of clinically relevant threshold probabilities (**Figure 6B**), demonstrating the clinical utility of the model.

We used SHAP in order to explain the biological processes that justify the predictions given by the "black-box" XGBoost model. The features were ranked according to their global effect on the model output in the SHAP summary plot (**Figure 7**). In keeping with its protective value, negative SHAP values were associated with high NLR and positive SHAP values with lower NLR. Also, the positive SHAP values were linked to a history of prior radiotherapy and increased BMI, which proved them to be risk enhancers. The SHAP dependence plots also indicated non-linear thresholds; the strongest effect of NLR was indeed most pronounced in cases where the values reached above 4.0, which a linear model has not been able to show in full. Furthermore, the SHAP summary plot explicitly delineated the clinical significance of other key predictors. Higher baseline BMI (represented by red dots) consistently yielded positive SHAP values, acting as a prominent risk enhancer. Biologically, excess adipose tissue functions as an active endocrine organ secreting pro-inflammatory adipokines, which may lower the threshold for systemic immune hyper-reactivity. Similarly, the presence of prior radiotherapy (Prior_RT) shifted SHAP values positively; this aligns with the biological premise that localized radiation induces neoantigen release and tissue inflammation, potentially priming the immune system and synergistically amplifying systemic immunotoxicity when followed by ICI therapy. Other categorical factors, including combination ther-

Table 5. Comprehensive performance metrics of prediction models in the independent testing set

Model	AUC (95% CI)	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	F1-score	+LR	-LR
XGBoost	0.871 (0.805-0.937)	82.0	81.6	82.3	73.8	87.9	0.782	4.61	0.22
Random Forest	0.849 (0.778-0.920)	79.0	76.3	80.6	70.7	84.7	0.734	3.93	0.29
Logistic Regression	0.795 (0.712-0.878)	75.0	71.1	77.4	65.9	81.4	0.684	3.15	0.37

Notes: Metrics were calculated at the threshold that maximized the Youden Index in the training set and applied to the testing set. Abbreviations: AUC, Area Under the Receiver Operating Characteristic Curve; CI, Confidence Interval; PPV, Positive Predictive Value; NPV, Negative Predictive Value; +LR, Positive Likelihood Ratio; -LR, Negative Likelihood Ratio. Note that +LR and -LR are used to denote likelihood ratios to avoid abbreviation conflict with the Neutrophil-to-Lymphocyte Ratio (NLR).

Table 6. Subgroup analysis of XGBoost model performance in the independent testing set

Subgroup	No. of Patients	No. of Events	AUC (95% CI)
Overall Cohort	100	28	0.871 (0.805-0.937)
Age			
≤ 65 years	56	16	0.882 (0.794-0.970)
> 65 years	44	12	0.854 (0.742-0.966)
Sex			
Male	67	19	0.868 (0.785-0.951)
Female	33	9	0.875 (0.745-1.000)
Histology			
Non-squamous	70	20	0.874 (0.792-0.956)
Squamous	30	8	0.865 (0.710-1.000)
PD-L1 TPS			
< 1%	34	10	0.880 (0.765-0.995)
≥ 1%	66	18	0.866 (0.780-0.952)
ICI Agent Type			
PD-1 Inhibitor	80	23	0.869 (0.782-0.956)
PD-L1 Inhibitor	20	5	0.876 (0.695-1.000)
Treatment Regimen			
Monotherapy	39	9	0.885 (0.760-1.000)
Combination Therapy	61	19	0.862 (0.771-0.953)

Notes: 95% CI was calculated using 2000 bootstrap replicates. The expanded analysis confirms the model's robust discriminative capability across key therapeutic variables. Abbreviations: AUC, Area Under the Curve; CI, Confidence Interval; ICI, Immune Checkpoint Inhibitor; PD-1, Programmed Cell Death 1; PD-L1, Programmed Death-Ligand 1; TPS, Tumor Proportion Score; XGBoost, Extreme Gradient Boosting.

apy regimens and baseline PPI use, also demonstrated rightward shifts in SHAP values, reflecting their respective additive roles in perturbing immune homeostasis and altering the gut microbiome.

Lastly, to allow the usability of our findings at the bedside, we discussed the trade-off between predictive precision and usability. Despite the fact that the XGBoost model had better discrimination (AUC = 0.871), the logistic regression model had reasonable discrimination (AUC = 0.795) and inherent transparency. To create a parsimonious clinical tool, a static nomogram was constructed based on a refined multivariable logistic regression model retaining only predictors with $P < 0.10$ (**Figure 8**). This visual tool is effectively a combination of the weighted values of the top 6 most significant predictors: Prior Radiotherapy, Combination Therapy, PPI use, BMI, NLR, and TSH, and enables clinicians to quickly estimate a specific probability of cs-irAEs. To illustrate this, a patient with a total score of 120 points would have a probability of about 60% of developing irAE. Crucially, to ensure the reliability of this bedside tool, we rigorously validated the nomogram within the independent testing set. The simplified nomogram demonstrated excellent calibration, with the predicted probabilities aligning closely with the observed frequencies (Hosmer-Lemeshow test, $P = 0.312$). Furthermore, DCA confirmed that employing this nomogram yielded a positive clinical net benefit across a practical threshold probability range of 5% to 45%. Clinically, this threshold probability represents the specific risk level at which a physician would decide that the expected benefit of implementing intensive monitoring outweighs the potential burden. The positive net benefit within this 5%

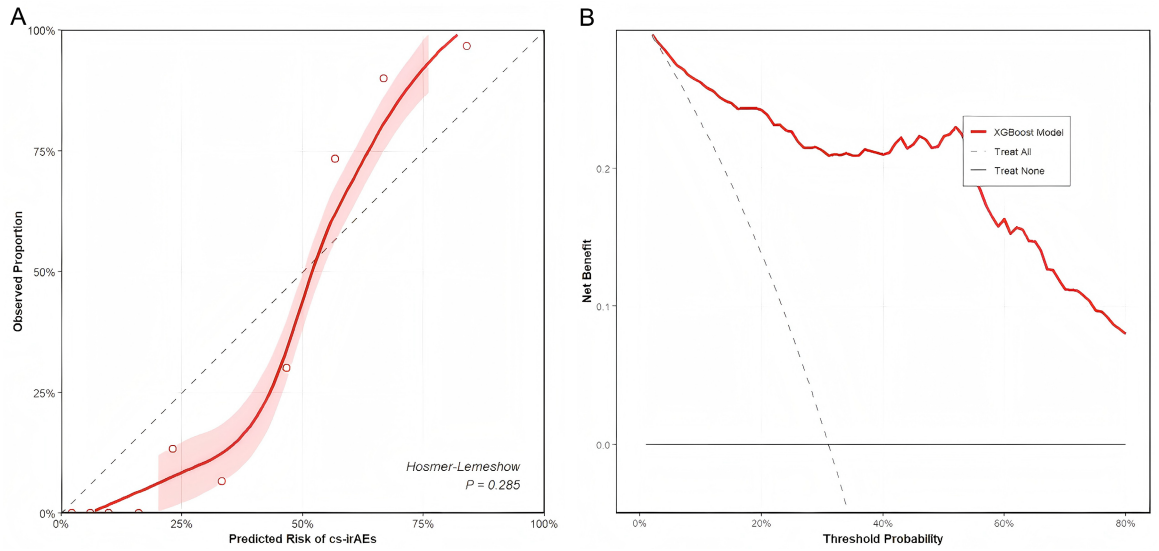

Figure 6. Calibration and decision curve analysis of the optimal XGBoost model. A. Calibration curve of the XGBoost model in the independent testing set. B. DCA estimates the clinical net benefit. The model shows positive net benefit across the threshold probability range of 10%-70%. Abbreviations: DCA, Decision Curve Analysis; cs-irAEs, Clinically Significant Immune-Related Adverse Events; XGBoost, Extreme Gradient Boosting.

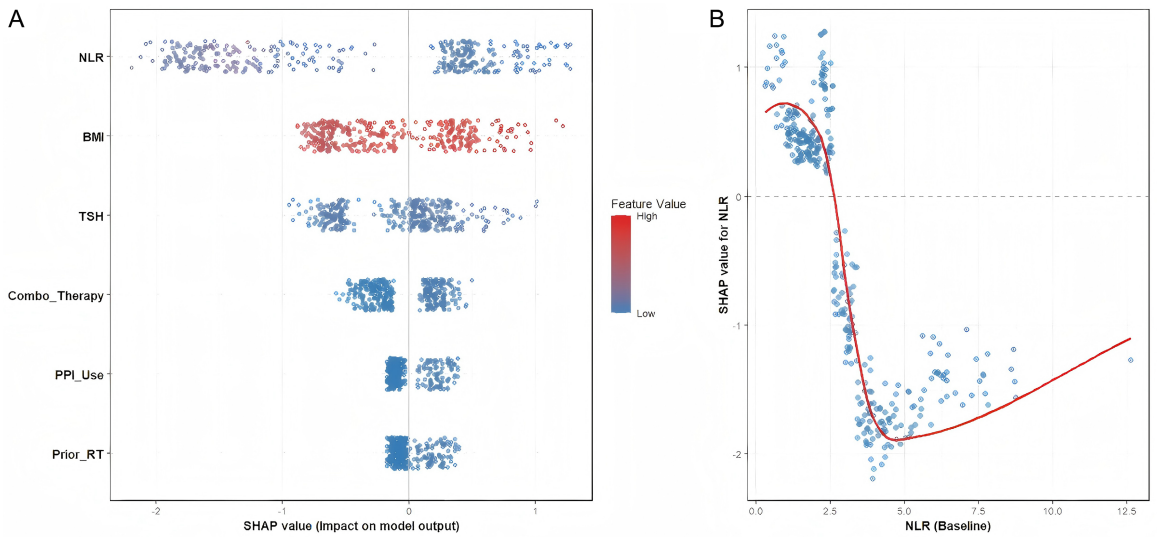

Figure 7. Model interpretation via SHAP. A. SHAP summary plot for the XGBoost model. Each dot represents a patient. The position on the x-axis indicates the SHAP value (impact on model output): positive values indicate increased risk, while negative values indicate decreased risk. The color indicates the feature value (red = high, blue = low). As shown, high NLR (red) is associated with negative SHAP values (lower risk). B. SHAP dependence plot for NLR, illustrating the non-linear relationship between NLR levels and the associated risk impact. Abbreviations: BMI, Body Mass Index; cs-irAEs, Clinically Significant Immune-Related Adverse Events; NLR, Neutrophil-to-Lymphocyte Ratio; PPI, Proton Pump Inhibitor; RT, Radiotherapy; SHAP, SHapley Additive exPlanations; TSH, Thyroid-Stimulating Hormone; XGBoost, Extreme Gradient Boosting.

to 45% range indicates that, for any clinician whose decision threshold falls within this interval, utilizing the nomogram to guide monitoring strategies results in superior patient outcomes - effectively identifying more true high-risk patients without imposing an excessive burden of false positives - compared to the default empirical strategies of monitoring all patients

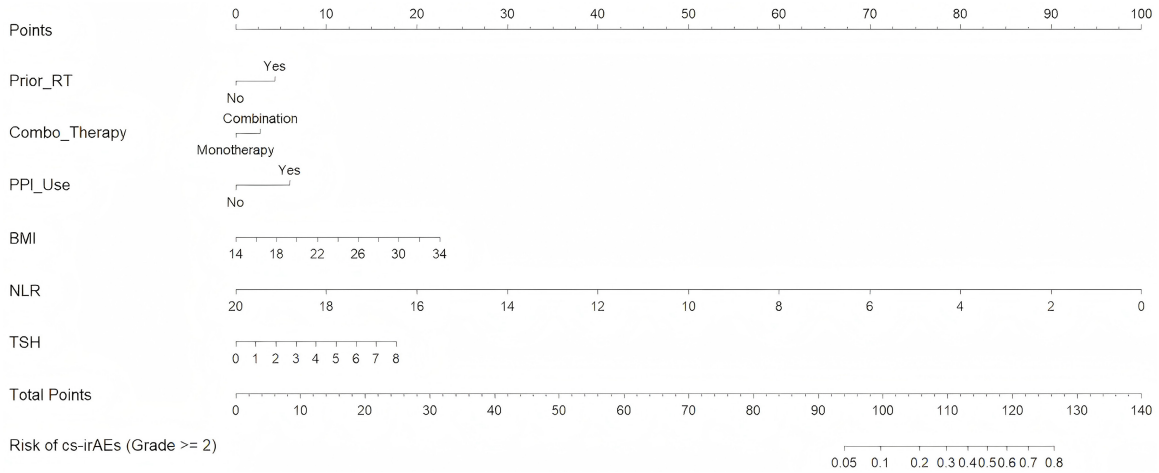

Figure 8. Clinical nomogram for predicting cs-irAEs. A static nomogram was constructed based on a refined multivariable logistic regression model incorporating six independent predictors (Prior Radiotherapy, Combination Therapy, PPI use, BMI, NLR, and TSH). Usage instructions: To calculate the individual risk, locate the patient's value for each predictor variable on its respective axis, and draw a vertical line upward to the "Points" axis to determine the assigned score for that specific variable. Sum the scores for all six predictors to obtain the "Total Points". Finally, locate this sum on the "Total Points" axis and draw a vertical line downward to the "Risk of cs-irAEs" axis to determine the patient's predicted probability of developing clinically significant immunotoxicity. Abbreviations: BMI, Body Mass Index; cs-irAEs, Clinically Significant Immune-Related Adverse Events; NLR, Neutrophil-to-Lymphocyte Ratio; PPI, Proton Pump Inhibitor; RT, Radiotherapy; TSH, Thyroid-Stimulating Hormone.

or monitoring none. The calibration curve and DCA for the nomogram have been provided in Figure S2, substantiating its safety and validity for real-world clinical application.

Discussion

The proposed study leveraged a gap in the literature of advanced NSCLC management by developing and validating a machine learning-based framework to predict cs-irAEs. Our main conclusion is that an XGBoost model that incorporates a wide range of clinical and laboratory data exhibits outstanding predictive power with an AUC of 0.871 in an independent test, significantly outperforming traditional logistic regression models. We identified several important independent predictors of cs-irAEs, including a history of prior radiotherapy, combination immunotherapy regimens, concomitant use of PPIs, high BMI, and a paradoxically beneficial effect of high NLR. To translate these findings into clinical practice, we also created a convenient nomogram for individual risks stratification.

Our XGBoost model (AUC = 0.871) performs better than numerous published irAE prediction models applied in NSCLC that have typically had AUCs between 0.75 and 0.85 [19, 20]. Several of these previous models had a drawback because they relied on traditional logistic regression, which might not accurately represent the non-linear interactions between predictors and irAE risk [21]. In addition, much of the previous research centered on a few biomarkers or clinical variables. We were instead able to take a holistic approach to our study, where we considered a large number of variables and used the power of machine learning to untangle their complex interrelationships. One of the conceptual advances of our study is the application of the Fine-Gray competing risk model that considers the high mortality rate of patients with advanced NSCLC, which may preclude the occurrence of irAEs - an essential element that has mostly been ignored in earlier predictive models [22, 23]. Our model provides a more precise estimate of the real rate of irAE occurrence, thereby making it a more clinically useful and reliable tool by accounting for this competing risk.

The observation that previous radiotherapy is an independent risk factor for cs-irAEs is consistent with the existing knowledge on the systemic immunological consequences of radiation. This is thought to be responsible for the

so-called abscopal effect, in which localized radiation causes immunogenic cell death and results in the release of tumor-associated antigens and damage-associated molecular patterns [24, 25]. This, subsequently, stimulates antigen-presenting cells and primes a systemic, T-cell mediated anti-tumor immune response. Nevertheless, an increased degree of immune activation may also disturb the tolerance to self-antigens, resulting in off-target immune attack and the occurrence of irAEs [26]. Likewise, it is likely that the correlation between combination therapy and heightened risk of irAE is attributable to the immunomodulatory properties of cytotoxic agents. Numerous chemotherapeutic agents have the ability to selectively deplete immunosuppressive cell populations, most commonly regulatory T cells, which in combination with ICIs lower the immune system setpoint and predispose individuals to autoimmune disease onset [27].

The results of our work demonstrated a rather counter-intuitive negative correlation between NLR and risk of irAEs, a finding that has been reported inconsistently but is the focus of discussion [28]. Our hypothesis is that a high level of baseline NLR, although it is a well-established negative prognostic factor for cancer outcomes, is a manifestation of systemic inflammation and relative lymphopenia, which can be a sign of underlying immune suppression or exhaustion [29]. Such an environment can cause the immune system to be less able to respond vigorously not only to the tumor but also to self-antigens, thereby leading to a reduced rate of irAEs. Crucially, our study's SHAP dependence analysis provided granular empirical data to refine this hypothesis, demonstrating a stark, non-linear threshold effect. As visualized in our SHAP plots, the protective impact against cs-irAEs becomes uniquely pronounced only when baseline NLR exceeds 4.0. Biologically, we postulate that this specific 4.0 threshold represents a critical tipping point of systemic immune exhaustion; beyond this level, the severely depleted peripheral lymphocyte reserves and the dominant neutrophil-driven suppression simply cannot mount the robust, hyper-reactive autoimmune cascade required to trigger significant irAEs [30, 31].

The discovery of PPI use as an independent predictor of cs-irAEs contributes to the accumulated evidence that the gut microbiome is involved in controlling systemic immunity and long-term outcomes of immunotherapy [32]. While our retrospective study design inherently precludes direct experimental verification through gut microbiota sequencing, our mechanistic hypothesis is strongly supported by recent, direct clinical literature and corroborated by our own model's output. Within our cohort, the SHAP summary analysis unequivocally demonstrated that baseline PPI use consistently shifted the predictive output toward a higher irAE risk profile. Recent robust cohort studies have explicitly demonstrated that chronic PPI use profoundly alters gut microbiota integrity (e.g., inducing severe hypochlorhydria-related dysbiosis), directly predisposing patients to more severe irAEs, particularly gastrointestinal toxicities, during ICI therapy [33, 34]. The changes in gastric pH induced by PPIs may lead to considerable alterations in the composition of the gut microbiota (dysbiosis). This may lead to a decline in beneficial bacteria that play an important role in maintaining immune homeostasis, such as *Akkermansia muciniphila* and *Faecalibacterium prausnitzii*. The healthy gut microbiome is crucial in balancing the systemic immune system as well as immune tolerance [34]. These functions may be disrupted by PPI-induced dysbiosis, which can decrease the threshold for autoreactivity and predispose patients to irAEs - a biological cascade that perfectly mirrors the elevated risk scores assigned to PPI users by our XGBoost algorithm. The presented finding highlights the need to use PPI judiciously in patients receiving immunotherapy and provides a rationale for investigating microbiome-modulating solutions to curtail the risk of irAEs in the future.

Our research possesses a number of methodologic strengths which enhance the validity and generalizability of our results. One of the main strengths was the use of the Fine-Gray competing risk model, as this captures all-cause mortality as a competing event, thus providing a more precise estimate of the incidence of irAEs in this high-risk group. Secondly, we used a holistic LASSO approach for feature selection that involves all variables simultaneously, thereby avoiding the biases inherent in univariate pre-selection strategies and enabling the identification of a more robust and clinically significant set of predictors. Moreover, our two-

model system (as a synthesis of the excellent predictive power of the XGBoost model and the clinical usefulness of a simplified nomogram based on logistic regression) can be considered a viable solution to the problem of integrating complex machine learning models into clinical practice. Lastly, the fact that our model was tested on an independent test set and demonstrated high performance provides strong evidence that our model is generalizable.

However, this study was limited in a number of ways. To begin with, its retrospective and single-center design made it vulnerable to selection bias and confounding, while also inherently limited the external applicability and generalizability of our model to broader, more diverse clinical populations. Consequently, rigorous external validation in multi-center, prospective cohorts is essential. Second, due to the limited sample size of specific outcome events, our model primarily predicted the overall occurrence of cs-irAEs and failed to capture or independently stratify the risk for rare, organ-specific irAE types (e.g., myocarditis or severe neurological toxicities), which may possess distinct pathophysiologic mechanisms. Third, we described PPI use as an exposure and postulated that it acted through the gut microbiome, but did not have direct metagenomic or metabolomic data to support our hypothesis. Future prospective studies should incorporate microbiome analysis to confirm this finding and explain the underlying mechanisms. Finally, the scope of our research was focused on irAE risk prediction and did not include efficacy data, such as progression-free survival and overall survival. The next step is critical risk-benefit analysis, which balances the risk of toxicity with therapeutic benefit. Future studies should focus on models that integrate efficacy and toxicity prediction, which would be a step further in terms of immunotherapy personalization.

Conclusion

This study developed a multidimensional machine learning-based model that demonstrated good predictive performance for estimating the risk of cs-irAEs within our study cohort of advanced NSCLC patients receiving PD-1/PD-L1 inhibitors. We found various easily accessible clinical and laboratory parameters, such as previous radiotherapy, combination therapy, PPI use, BMI, and NLR, as crucial predictors of irAEs. The nomogram created during this study is a simple and convenient tool to help clinicians recognize high-risk patients in order to provide them with individualized monitoring and preventive management options. However, considering the current clinical application thresholds, this predictive framework requires further external validation in larger, multi-center prospective cohorts before widespread implementation in routine practice. Ultimately, this tool provides a valuable foundation for reducing the severe effects of irAEs and delivering safer, more personalized immunotherapy to a broader patient population.

Disclosure of conflict of interest

None.

Address correspondence to: Mao Huang, Department of Respiratory and Critical Care Medicine, The First Affiliated Hospital with Nanjing Medical University, Nanjing 210000, Jiangsu, China. E-mail: mhuang828@163.com

References

[1] Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I and Jemal A. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2024; 74: 229-263.

[2] Gandhi L, Rodriguez-Abreu D, Gadgeel S, Esteban E, Felip E, De Angelis F, Domine M, Clingan P, Hochmair MJ, Powell SF, Cheng SY, Bischoff HG, Peled N, Grossi F, Jennens RR, Reck M, Hui R, Garon EB, Boyer M, Rubio-Viqueira B, Novello S, Kurata T, Gray JE, Vida J, Wei Z, Yang J, Raftopoulos H, Pietanza MC and Garassino MC. Pembrolizumab plus chemotherapy in metastatic non-small-cell lung cancer. N Engl J Med 2018; 378: 2078-2092.

[3] Jayathilaka B, Kader M and Thilakarathne NN. Cancer and treatment specific incidence rates of immune-related adverse events, a systematic review and meta-analysis. Sci Rep 2024; 14: 2298.

[4] Haanen J, Obeid M, Spain L, Carbonnel F, Wang Y, Robert C, Lyon AR, Spano JP, Treister S, Peccatori F, McElwee K, Ascierto PA, Peters S, Reck M, Ferte C, Bhatia S, Lorigan P, Blank CU, Litiere S, Reni M and Dougan M. Management of toxicities from immunotherapy: ESMO Clinical Practice Guideline for diagnosis, treat-

ment and follow-up. Ann Oncol 2022; 33: 1217-1238.
[5] Liang Y, Wang Z, Zhang J and Wang L. Biomarkers for immune-related adverse events in cancer immunotherapy. Jpn J Clin Oncol 2024; 54: 365-373.
[6] Nardini C. Machine learning in oncology: a review. Ecancermedicalscience 2020; 14: 1065.
[7] Goldstraw P, Chansky K, Crowley J, Rami-Porta R, Asamura H, Eberhardt WE, Nicholson AG, Groome P, Mitchell A and Bolejack V. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (Eighth) edition of the TNM classification for lung cancer. J Thorac Oncol 2016; 11: 39-51.
[8] Schneider BJ, Naidoo J, Santomasso BD, Lacchetti C, Adkins S, Anadkat M, Atkins MB, Brassil KJ, Caterino JM, Chau I, Davies MJ, Ernstoff MS, Fecher L, Ghosh M, Hwang I, Judson CH, Kennedy LC, MacAlister W, McGuire A, Meyerson C, Nguyen J, Puzanov I, Shirai K, Urba W, Vidal J, Wang Y, Wender RC, Wood R and Brahmer JR. Management of immune-related adverse events in patients treated with immune checkpoint inhibitor therapy: ASCO guideline update. J Clin Oncol 2021; 39: 4073-4126.
[9] Peduzzi P, Concato J, Kemper E, Holford TR and Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49: 1373-1379.
[10] Zhou X, Yao Z, Yang H, Liang N, Zhang X and Zhang F. Are immune-related adverse events associated with the efficacy of immune checkpoint inhibitors in patients with cancer? A meta-analysis. BMC Med 2020; 18: 87.
[11] US Department of Health and Human Services. Common terminology criteria for adverse events (CTCAE) version 5.0. US Department of Health and Human Services 2017.
[12] Oken MM, Creech RH, Tormey DC, Horton J, Davis TE, McFadden ET and Carbone PP. Toxicity and response criteria of the Eastern Cooperative Oncology Group. Am J Clin Oncol 1982; 5: 649-655.
[13] Zhao S, Gao G, Li W, Li X, Zhao C, Jiang T, Jia Y, He Y, Li A, Su C, Zhang J and Zhou C. Antibiotics are associated with attenuated efficacy of anti-PD-1/PD-L1 therapies in Chinese patients with advanced non-small cell lung cancer. Lung Cancer 2019; 130: 10-17.
[14] Mezquita L, Auclin E, Ferrara R, Charrier M, Remon J, Planchard D, Ponce S, Areses MC, Leroy L, Audigier-Valette C, Felip E, Bria E, Bironzo P, Areses C, Campisi C, Mazieres J, Hureaux J, Morere JF, Menis J, Guisier F, Leduc C, Swalduz A, Gounant V and Besse B. Association of the lung immune prognostic index with immune checkpoint inhibitor outcomes in patients with advanced non-small cell lung cancer. JAMA Oncol 2018; 4: 351-357.
[15] White IR, Royston P and Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011; 30: 377-399.
[16] Chawla NV, Bowyer KW, Hall LO and Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res 2002; 16: 321-357.
[17] Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Methodol 1996; 58: 267-288.
[18] Lundberg SM and Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017; 30.
[19] Gao W, Liu Q, Zhou Y, Wang X, Zhang Y, Li Y and Chen X. The predictive model construction for immune-related adverse events in non-small cell lung cancer patients receiving immunotherapy. Technol Cancer Res Treat 2023; 22: 15330338231206705.
[20] Gong L, Gong J, Sun X, Wang Y, Zhang X, Li Z and Chen H. Identification and prediction of immune checkpoint inhibitors-related pneumonitis by machine learning. Front Immunol 2023; 14: 1138489.
[21] Zhang Z, Cortese G, Combes F, Pellino G, Yoon WK, Ahmed AM, Nason G and Dong W. Statistical models versus machine learning for competing risks analysis. BMC Med Res Methodol 2023; 23: 38.
[22] Fu J, Zhang Z, Zhou L, Wang Y, Li M, Chen Y and Liu X. Competing risk and random survival forest models for predicting the benefit of adjuvant chemotherapy in stage II colon cancer. Sci Rep 2025; 15: 5824.
[23] Luo Y, Wang P, Wang K, Zhang Y, Li X, Zhao M and Chen L. Multiparameter prediction model of immune checkpoint inhibitors in patients with advanced solid tumors based on competing risk model. Sci Rep 2023; 13: 4059.
[24] Brooks ED and Chang JY. Abscopal effect of radiotherapy combined with immune checkpoint inhibitors. J Thorac Dis 2019; 11: S94-S99.
[25] Dai C, Liu F and Lu X. Abscopal effect: from a rare phenomenon to a new frontier in cancer therapy. Radiat Oncol 2024; 19: 77.
[26] Mondini M, Levy A, Meziani L, Milliat F, Deutsch E and Castera L. Radiation-induced bystander and abscopal effects: important lessons from the past, and new challenges for the future. Br J Cancer 2020; 123: 1083-1095.
[27] Zheng H, Zeltsman M, Zauderer MG, Eguchi T, Perez-Rojas A, Moussaly E, Rizk N, Jones DR and Adusumilli PS. Chemotherapy-induced immunomodulation in non-small-cell lung can-

cer: a rationale for combination chemoimmunotherapy. Immunotherapy 2017; 9: 913-927.
[28] Rossi T, Ziaco F, Tomao F, Di Lisa FS, Giorgi G, Papa A and Tomao S. The biomarkers related to immune related adverse events caused by immune checkpoint inhibitors in cancer treatment. Cancer Cell Int 2022; 22: 163.
[29] Ray A and Das A. Immune-related adverse events and the balancing act of immunotherapy. Nat Commun 2022; 13: 721.
[30] Crea F and Ciamporcero E. The obesity paradox in cancer, tumor immunology, and immunotherapy. Front Immunol 2019; 10: 1940.
[31] Vadevoo S, Schuler K, Brune B and Weigert A. The ambiguous role of obesity in oncology by promoting cancer but enhancing response to therapy. Cancer Metab 2022; 10: 4.
[32] Li H, Wu Q, Liu B, Zhang Y, Wang X, Chen Y and Li Z. Gut microbiota shapes cancer immunotherapy responses. NPJ Biofilms Microbiomes 2025; 11: 38.
[33] Shatila M, Devalaraju S, Takigawa K, Wang Y, Chen X, Li Z and Zhang Y. Worse survival and gastrointestinal toxicity outcomes among patients receiving proton pump inhibitors during checkpoint inhibitor therapy. J Natl Compr Canc Netw 2025; 23: e257023.
[34] Lasagna A, Mascaro F, Figini S, Lenti MV, Ferraris E, Corbella M, Baldanti F, Pedrazzoli P and Di Sabatino A. Impact of proton pump inhibitors on the onset of gastrointestinal immune-related adverse events during immunotherapy. Cancer Med 2023; 12: 19530-19536.

Table S1. Missing rates and consistency evaluation of variables before and after MICE imputation (N = 332)

Variable	Missing Count (Rate)	Original Data (Before Imputation)	Pooled Imputed Data (After Imputation)	Statistic (t or Z value)	*P*-value
TSH (mIU/L)	52 (15.6%)	2.15 [1.38, 3.22]	2.14 [1.40, 3.19]	Z = 0.145	0.884
CRP (mg/L)	41 (12.5%)	8.80 [3.35, 18.70]	8.75 [3.40, 18.60]	Z = 0.092	0.926
Albumin (g/L)	18 (5.4%)	38.65 ± 4.30	38.62 ± 4.28	t = 0.091	0.927
BMI (kg/m^2)	11 (3.2%)	23.31 ± 3.29	23.32 ± 3.28	t = -0.039	0.968

Notes: Data are presented as Mean ± SD for normally distributed variables (Albumin, BMI) and evaluated using the independent samples t-test; Data are presented as Median [IQR] for non-normally distributed variables (TSH, CRP) and evaluated using the Mann-Whitney U test. P > 0.05 indicates no significant difference between the original and imputed datasets. Abbreviations: BMI, Body Mass Index; CRP, C-Reactive Protein; IQR, Interquartile Range; MICE, Multiple Imputation by Chained Equations; SD, Standard Deviation; TSH, Thyroid Stimulating Hormone.

Table S2. Comparison of the univariate Fine-Gray sub-distribution hazard model and the standard Cox proportional hazards model for significant baseline predictors (N = 332)

Variable	Fine-Gray Model (Competing Risk Adjusted)		Standard Cox Model (Unadjusted for Competing Risk)	
	SHR (95% CI)	*P*-value	HR (95% CI)	*P*-value
BMI (per 1 kg/m^2 increase)	1.08 (1.02-1.14)	0.009	1.11 (1.04-1.18)	0.003
Therapy Type (Combination vs. Mono)	1.55 (1.02-2.36)	0.041	1.68 (1.09-2.58)	0.018
Prior Radiotherapy (Yes vs. No)	1.62 (1.05-2.48)	0.029	1.74 (1.11-2.72)	0.015
NLR (per 1-unit increase)	0.85 (0.76-0.95)	0.005	0.81 (0.71-0.92)	0.001
dNLR (per 1-unit increase)	0.82 (0.71-0.94)	0.006	0.77 (0.65-0.90)	0.001

Notes: This sensitivity analysis compares the effect sizes generated by the Fine-Gray competing risk regression against the standard Cox proportional hazards regression. The standard Cox model treats the competing event (all-cause death prior to the onset of cs-irAEs) as independent, right-censored data. As demonstrated, this traditional approach consistently overestimates the magnitude of the risk associations (yielding hazard ratios further from the null value of 1.0) compared to the competing-risk adjusted Fine-Gray model, thereby justifying the use of the latter for our primary analysis. Abbreviations: BMI, Body Mass Index; CI, Confidence Interval; cs-irAEs, Clinically Significant Immune-Related Adverse Events; dNLR, derived Neutrophil-to-Lymphocyte Ratio; HR, Hazard Ratio; NLR, Neutrophil-to-Lymphocyte Ratio; SHR, Subdistribution Hazard Ratio.

Table S3. Median onset time and clinical outcomes of specific immune-related adverse events (N = 93 events)

irAE Organ System	Total Events (n)	Median Onset Time, months [IQR]	Outcome: Resolved/Clinically Controlled*, n (%)
Dermatitis (Rash/Pruritus)	33	1.8 [1.0-3.2]	31 (93.9%)
Endocrine (Thyroiditis)	22	2.5 [1.5-4.5]	22 (100.0%)*
Pneumonitis	12	3.6 [2.1-6.0]	9 (75.0%)
Colitis	9	2.0 [1.2-3.8]	8 (88.9%)
Hepatitis	8	2.8 [1.8-4.2]	6 (75.0%)
Nephritis	4	4.1 [2.5-6.5]	3 (75.0%)
Other (e.g., Myositis)	5	3.2 [2.0-5.5]	4 (80.0%)

Notes: Outcomes are defined as the resolution of the toxicity to Grade ≤ 1 or a return to baseline severity. *For endocrine toxicities (e.g., hypothyroidism following destructive thyroiditis), true physiologic resolution is rare. Therefore, "Clinically Controlled" denotes patients who remain asymptomatic and clinically stable on long-term physiologic hormone replacement therapy (e.g., levothyroxine). Abbreviations: IQR, Interquartile Range; irAEs, Immune-Related Adverse Events.

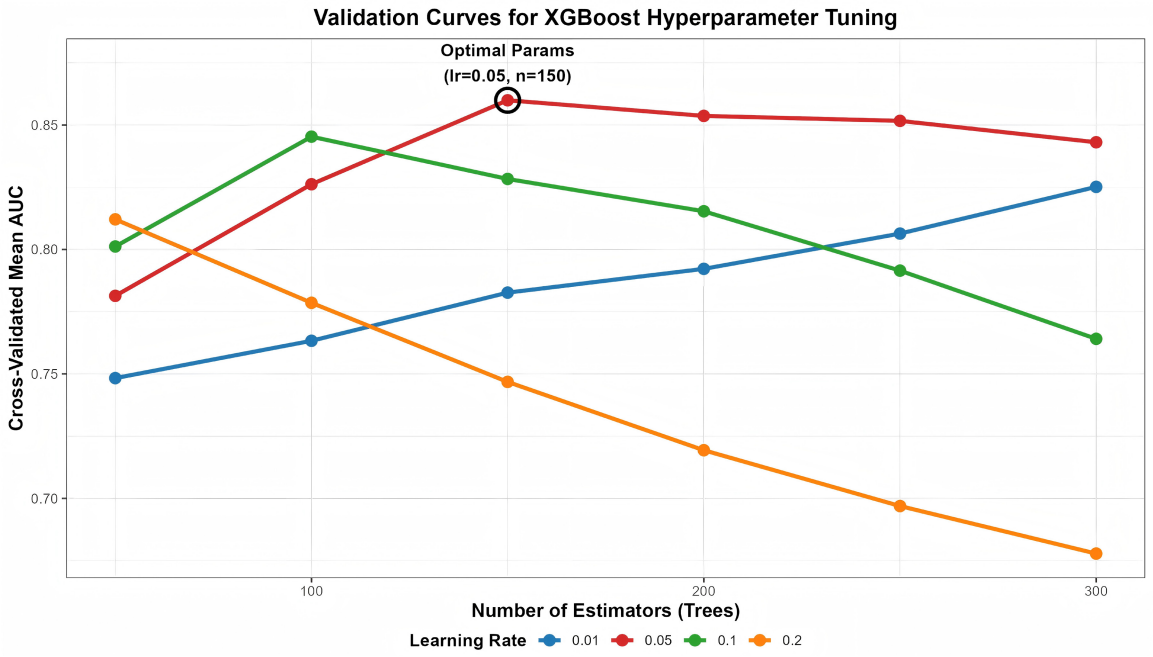

Figure S1. Validation curves for XGBoost model hyperparameter tuning. The plot illustrates the dynamic changes in the cross-validated mean AUC across a comprehensive grid search space. The x-axis represents the number of estimators (trees) ranging from 50 to 300, while the y-axis indicates the mean AUC derived from 5-fold cross-validation on the training set. Distinct colored lines correspond to different learning rates (0.01, 0.05, 0.1, and 0.2). The black open circle highlights the optimal hyperparameter combination (learning rate = 0.05, n_estimators = 150), which successfully maximizes the model's discriminative performance (mean AUC = 0.862) prior to the onset of overfitting typically observed with higher learning rates or excessive tree growth. Abbreviations: AUC, Area Under the Curve; XGBoost, Extreme Gradient Boosting.

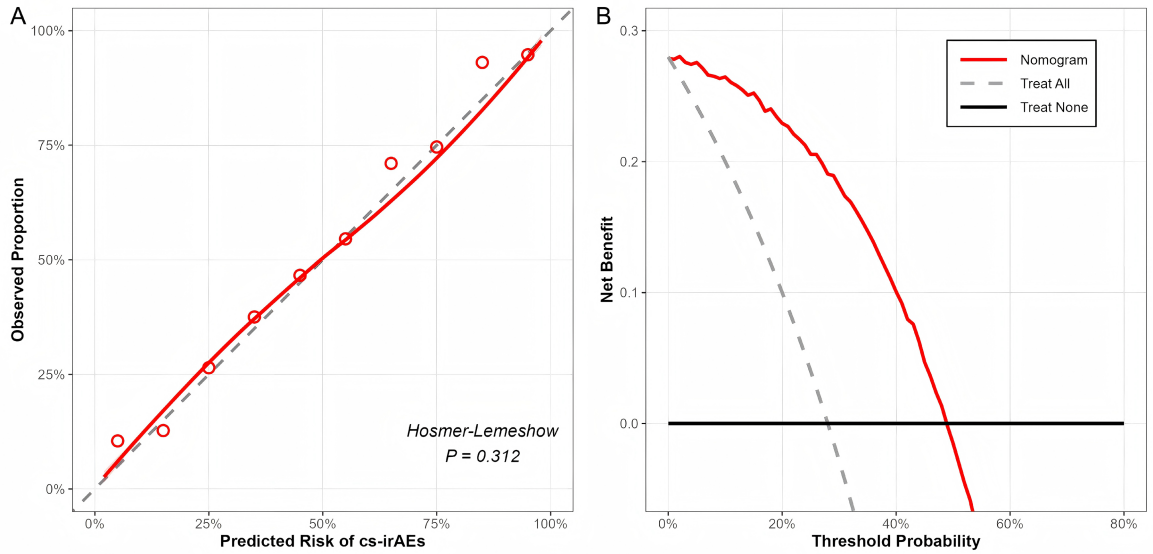

Figure S2. Internal validation of the constructed clinical nomogram in the independent testing set. A. Calibration curve demonstrating the agreement between the nomogram-predicted probabilities of cs-irAEs and the actual observed frequencies. The Hosmer-Lemeshow test resulted in a non-significant *P*-value of 0.312, indicating excellent model calibration without substantial deviation. B. DCA of the nomogram. The solid red line represents the net benefit of using the nomogram to guide clinical monitoring. Compared to the default strategies of monitoring all patients ("Treat All", dashed grey line) or monitoring none ("Treat None", solid black line), the nomogram consistently yields a positive clinical net benefit across a wide and practical threshold probability range spanning from approximately 5% to 45%. Abbreviations: DCA, Decision Curve Analysis; cs-irAEs, Clinically Significant Immune-Related Adverse Events.