Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Longitudinal Study on Predicting Quality of Life in T3 Stage Radiotherapy Patients Based on Ensemble Learning and SHAP Interpretation

Yanze Shi
Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital & Institute)
Jiayin Liu
Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital & Institute)
Yu Tian
Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital & Institute)
Wei Zheng
Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital & Institute)
Hui Li
lnzllihui@163.com
Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital & Institute)

Research Article

Keywords: Radiotherapy, Abdominal Malignancies, Quality of Life, Ensemble Learning, Stacking Model, SHAP Analysis, Predictive Model, Longitudinal Study, Dynamic Feature Engineering, Clinical Threshold, Precision Oncology

Posted Date: July 13th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10008451/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

A Longitudinal Study on Predicting Quality of Life in T3 Stage Radiotherapy Patients Based on Ensemble Learning and SHAP Interpretation

Yanze Shi[1†], Jiayin Liu[1†], Yu Tian[1], Wei Zheng[1], Hui Li[1*]
[1]Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital & Institute), Shenyang 110042, Liaoning Province, P. R. China
[†]These authors contributed equally to this work.
[*]Corresponding author: Hui Li, Email:lnzllihui@163.com

Abstract

Objective To accurately predict the quality of life (QoL) of patients with abdominal malignant tumors three months after radiotherapy (T3 phase), identify the core modifiable determinants of the quality of life at this stage, and determine their quantified clinical thresholds, thereby providing evidence for precision oncology interventions and improving patient-centered outcomes.

Methods This single-center, prospective longitudinal observational study, includes 133 patients with abdominal malignancies undergoing radiotherapy from January 2019 to December 2021. Multidimensional clinical and scale data were collected at four time points: before radiotherapy (T0), midway through radiotherapy (T1, when the total radiotherapy dose reached 50%), at the end of radiotherapy (T2), and three months after radiotherapy (T3). A comprehensive analysis framework was developed, encompassing data preprocessing, feature engineering, ensemble modeling, and interpretability analysis. Following targeted data preprocessing and dynamic feature generation, a multidimensional validation system (classification performance, calibration performance, clinical utility, stability, etc.) was used to evaluate the model's performance. Combined with SHAP analysis, we explored core influencing factors, clinical thresholds, and feature interaction effects, and performed tumor subtype subgroup analyses to verify model stability.

Results The ensemble model performed excellently on the validation set, achieving an AUC of 0.929 (95% CI: 0.799-1.000). The accuracy, sensitivity, and specificity were 0.885, 0.923, and 0.846, respectively, and the calibration performance was good (Hosmer-Lemeshow test, P=0.151). Decision curve analysis confirmed that the model provided substantial net benefit within the core intervention threshold range of 0.10 – 0.80, with an absolute mean net benefit of 0.32. The number needed to treat (NNT) within this threshold range was 3. One hundred Monte Carlo simulations and multiple internal validations confirmed the model's excellent stability and robustness against methodological biases. Among the 133 patients, the incidence of low QoL at T3 was 39.09% (52/133), which is consistent with the sample size estimate. SHAP analysis clearly identified the Top-5 core factors affecting the quality of life at T3 as household chores ability (SHAP importance=0.0357, clinical threshold 0.263), age-standardized value (SHAP importance=0.0303, clinical threshold 0.511), age (SHAP importance=0.0298, clinical threshold 0.511), acceptance of physical changes after treatment

(SHAP importance=0.0268, clinical threshold 1.98) and the degree of sadness and distress (SHAP importance=0.0204, no clinical threshold), a synergistic interaction was found between household chores ability and age (SHAP interaction value=0.012, *P*=0.031); Subgroup analysis by tumor subtypes showed that the core influencing factors for gastric cancer patients were household chores ability and the degree of sadness and distress, while for colorectal cancer patients, they were age-standardized values and acceptance of physical changes after treatment. Moreover, the Top-5 core factors for each subtype were identical to the overall sample (overall sample Kappa=0.92, *P*<0.001).

Discussion The Stacking ensemble model constructed in this study demonstrates high predictive accuracy, good stability, and clinical practicality for the T3 phase quality of life in patients undergoing radiotherapy for abdominal malignancies. It provides a reproducible methodological framework for machine-learning analysis of small-sample longitudinal oncology data. The core influencing factors, quantifiable clinical thresholds, and feature interaction effects revealed by SHAP analysis offer key clinical hypotheses for future large-sample, multicenter studies predicting QoL in radiotherapy patients. They also establish core clinical targets for precise, individualized rehabilitation interventions aimed at T3 QoL in patients undergoing radiotherapy for abdominal malignant tumors, while providing scientific reference for formulating differentiated intervention strategies for different tumor subtypes. This study provides Level IIb evidence. The results are applicable to adult patients with abdominal malignant tumors receiving intensity-modulated radiotherapy (IMRT) with a total dose of 45-60 Gy. It can serve as a methodological reference for subsequent large-sample, multi-center studies and provide preliminary evidence for the accumulation of data in the quality of life management section of future guidelines for patients undergoing radiotherapy for abdominal malignant tumors.

Keywords Radiotherapy; Abdominal Malignancies; Quality of Life; Ensemble Learning; Stacking Model; SHAP Analysis; Predictive Model; Longitudinal Study; Dynamic Feature Engineering; Clinical Threshold; Precision Oncology

Introduction

Radiotherapy is a cornerstone of multimodal cancer treatment, playing an irreplaceable role in local tumor control and prolonging patient survival. However, the cumulative physical side effects of radiotherapy (e.g., radiation fibrosis, chronic fatigue, neuropathic pain, and radiation enteritis) and psychosocial distress (e.g., cancer-related post-traumatic stress disorder, body image disturbance, fear of recurrence, sadness, and depression) are long-lasting [1,2]. As a result, 40% – 60% of radiotherapy patients experience a significant decline in quality of life 3 to 6 months after treatment [3]. Among these, radiotherapy for abdominal malignancies is particularly challenging because the target volume often lies close to critical organs such as the gastrointestinal tract, peritoneum, and urinary system. Chronic side effects—including radiation enteritis, digestive dysfunction, and intra-abdominal adhesions—directly affect basic functions such as eating and defecation. These side effects directly affect patients' basic life

functions such as eating and defecating, and compared with head and neck or breast tumors, the negative impact on QoL is more persistent and specific. Currently, there is an international research gap regarding QoL prediction models specifically developed for this population In the era of precision medicine, early identification of patients at high risk of poor QoL and the implementation of personalized targeted interventions have become key scientific issues in improving overall cancer treatment outcomes and achieving "patient-centered" diagnosis and treatment [4].

QoL is a dynamic concept shaped by the interplay of physiological function, clinical characteristics, psychological state, and social support, and it continuously changes with the progression of the disease [5]. Traditional prediction models are often based on cross-sectional data at a single time point and built using a single algorithm. These models have inherent limitations, including limited predictive accuracy, poor generalizability, and an inability to capture dynamic disease trajectories [6]. In recent years, ensemble learning has shown outstanding performance in handling high-dimensional nonlinear medical data by integrating the strengths of multiple models. Longitudinal follow-up data can fully capture patients' disease progression trajectories, providing richer information for dynamic prediction. However, few studies have combined these two approaches to predicting the quality of life in radiotherapy patients. Additionally, the "black box" nature of complex machine learning models severely limits their clinical translation and practical application.

Although existing studies has explored the application of machine learning in predicting the quality of life of cancer patients, several limitations remain: domestic studies often use logistic regression or random forest alone, building models based on cross-sectional data, with AUCs mostly between 0.75 and 0.85; while international studies have attempted ensemble learning, they mostly focus on colorectal cancer as a single tumor type and have not yet covered abdominal malignancies such as gastric and liver cancers. Furthermore, the clinical thresholds have not been clearly quantified in conjunction with SHAP analysis, leaving a significant gap between the current state and the needs of precise clinical intervention. SHAP analysis, based on Shapley values from game theory, quantifies each feature's contribution to individual predictions, thereby clearly revealing the model' s predictive logic. It provides a powerful tool for addressing the "black box" problem of complex models and promoting clinical translation [10,11].

Currently, similar international studies mostly focus on single tumor types of colorectal cancer, and the model prediction AUC is mostly between 0.87 and 0.89 [7,21]. No longitudinal dynamic prediction model has yet been developed for multiple subtypes of abdominal malignancies, nor have clear, quantifiable clinical intervention thresholds been established. This study is based on longitudinal follow-up data from T0-T3 radiotherapy patients with abdominal malignancies, integrating longitudinal dynamic feature engineering with Stacking ensemble learning, and combining SHAP analysis to achieve model interpretability and clinical threshold quantification, aiming to address the shortcomings of existing research.

Accordingly, we constructed a comprehensive research framework encompassing data preprocessing, feature engineering, ensemble modeling, and interpretability analysis. The aims were to: ① build and validate a high-performance Stacking ensemble prediction model for T3 stage quality of life in patients with multiple subtypes of abdominal malignant tumors undergoing radiotherapy; ② identify the core modifiable factors affecting T3 stage quality of life in this population, and determine their quantified clinical thresholds through dual validation using SHAP importance and ROC Youden index; ③ confirm the synergistic interaction effects of features. This study focuses on highly homogeneous a highly homogeneous cohort of patients receiving radiotherapy for abdominal malignancies, providing a methodological exploration of predictive research combining small-sample longitudinal data with ensemble learning. The core findings will be further validated and extended in subsequent large-sample, multi-center studies.

Materials and Methods

Research Subjects and Data Sources

Inclusion and Exclusion Criteria

A total of 158 consecutive patients with abdominal malignancies who received intensity-modulated radiotherapy (IMRT) at our hospital from January 2019 to December 2021 were initially screened. After applying strict inclusion and exclusion criteria, 133 patients were finally included.

Inclusion criteria: ① diagnosis of gastric cancer, liver cancer, pancreatic cancer, gallbladder cancer, bladder cancer, or colorectal cancer confirmed by imaging or pathology; ② Abdominal malignant tumors with secondary metastasis; ③ A history of radiotherapy or currently undergoing radiotherapy; ④ Age ≥ 18 years; ⑤ Clear consciousness, able to communicate normally and cooperate with scale assessments; ⑥ Signed informed consent from patient and family members.

Exclusion criteria: ① Patients with metastasis or recurrence after comprehensive treatment; ② Patients with comorbid mental illnesses; ③ Patients with severe organ dysfunction; ④ Patients unaware of their own condition.

A total of 25 patients were excluded: 12 with metastatic recurrence, 5 with mental illness, and 8 with severe organ dysfunction (note: the excluded patients with metastatic recurrence had developed new metastases after radiotherapy, which distinguishes them from the initial diagnosis with secondary metastasis that was allowed by the inclusion criteria). All radiotherapy plans adhered to the NCCN Radiotherapy Version 2.2021 international clinical guidelines, using IMRT with a single fraction dose of 1.8–2.0 Gy and a total dose of 45–60 Gy [12].

Sample Size Estimation

Sample size was estimated using G*Power 3.1 software based on the Events Per Variable (EPV) rule (EPV ≥ 10). The expected incidence of low quality of life at T3 was approximately 40%. A total of 15 predictor variables were planned to be included. The minimum required sample size calculated was 15×10/0.4=125 cases. The actual number of cases included in this study was 133, meeting the statistical modeling requirements. For subtypes such as liver cancer and pancreatic cancer, because the sample size for each was <30, post-hoc power analysis showed statistical power <70%; therefore, stratified analysis was not performed. The core influencing factors for these subtypes will be further validated in subsequent multicenter studies.

Follow-up Time Points and Data Collection

Patients were followed up face-to-face at T0 (before radiotherapy), T1 (mid-radiotherapy, when the total dose reached 50%), T2 (end of radiotherapy), and T3 (three months after radiotherapy). Data were collected using standardized tools by oncology specialist nurses who had received uniform training. The test-retest reliability of all scales was ICC > 0.85, ensuring the consistency and accuracy of data collection. Data collection tools include:

1. General Information Questionnaire: covering two dimensions: sociodemographic information (age, gender, marital status, etc.) and clinical disease information (tumor type, stage, radiation dose, etc.);

2. Memorial Symptom Assessment Scale (MSAS-Ch, Chinese version): assesses the occurrence of 32 symptoms over the past week, with good reliability and validity [13];

3. Quality of Life Scale for Cancer Patients: physical, social/family, emotional, and functional. The total score ranges from 0 to 108, with higher scores indicating better quality of life [14];

4. Fear of Progression Questionnaire-Short Form (FoP-Q-SF, Chinese Version): Contains 2 dimensions and 12 items, with a total score ranging from 12 to 60. The higher the score, the greater the fear of disease progression [15];

5. Cancer Self-Management Scale (CSMS, Chinese version): contains 6 dimensions and 44 items, with higher scores indicating a higher level of self-management [16].

Initial Dataset Features

The initial dataset contained 532 samples (133 patients × 4 time points) and 149 features, categorized into four types: demographic variables (12), clinical variables (5), symptom variables (60), and scale variables (72). This study was approved by the Ethics Committee of our hospital (Approval No.: 2026-119). All patients signed written informed consent. The research process adhered to the principles of the Declaration of Helsinki, and clinical data were anonymized according to HIPAA guidelines, used solely for scientific analysis in this study.

Data Preprocessing

To improve data quality and eliminate confounding factors, data preprocessing performed in four steps: quality check - missing value handling - outlier handling - standardization. All steps were implemented using Python 3.9.7 to ensure that the data met the requirements for machine learning modeling.

1. Data Quality Check: Standardize feature names, remove line breaks, extra spaces, and special characters, verify the consistency of fields across time points T0-T3, eliminate missing/duplicate feature names; convert continuous and ordinal categorical non-identifying features to numerical type, and keep binary features in their original 0/1 encoding.

2. Handling Missing Values: Based on the Marchenko-Pastur distribution test, missing values in the dataset were confirmed to be missing at random (MAR, P=0.32). A time-point and variable-type specific strategy was used for imputation: for demographic variables with missing values at T0, the mean of same-period patients was used; for T1–T3, forward filling within the same patient was applied; for clinical variables, the median of same-period patients was used; and for symptom/scale variables, the mean of similar patients at the same time point was used. Sensitivity analysis showed that when multiple imputation (MI) and K-nearest neighbors (KNN, k = 5) were used instead of the original method, the model AUC fluctuated by less than 3%, and the top-five core features remained stable, confirming the validity of the imputation methods.

3. Outlier Handling: Use box plots to identify extreme values of numerical variables. For the five core clinical variables—age, height, weight, BMI, and total number of radiotherapy fractions—the capping method was applied according to the reference ranges set by the Chinese Clinical Oncology Diagnosis and Treatment Guidelines and related clinical studies. Extreme values that contradicted medical common sense, such as BMI=5, are directly were removed, after which all outliers for the five core variables were being set to zero (Table 1). The variance inflation factor (*VIF*) was also used to test for feature collinearity. A $VIF < 10$ indicated no significant collinearity, ensuring that the feature quality met modeling requirements.

Table 1. Results of Outlier Handling for Clinical Variables

Clinical Variable	Number of Outliers Before Processing	Number of Outliers After Processing	Truncation Range	Clinical Reference Range
Age	4	0	[27.50, 79.50]	[18, 100]
Height (m)	4	0	[1.46, 1.91]	[1.4, 2.2]
Weight (kg)	12	0	[30.00, 100.00]	[30, 200]
BMI	13	0	[15.00, 32.43]	[15, 50]
Total Number of	133	0	[25.00, 30.00]	[0, 100]

Clinical Variable	Number of Outliers Before Processing	Number of Outliers After Processing	Truncation Range	Clinical Reference Range
Radiotherapy Fractions				

Note: Box plots were used to identify extreme values of numerical variables. Combined with the reference ranges set by the "Chinese Clinical Oncology Diagnosis and Treatment Guidelines" and related clinical studies, the upper and lower limit capping method was applied to five core clinical variables: age, height, weight, BMI, and total number of radiotherapy sessions, directly removing extreme values that contradict medical common sense.

4. Standardization: Continuous features such as age, BMI, and total scale scores were standardized using Z-score transformation, converting them into a standard normal distribution with a mean of approximately 0 and a standard deviation of approximately 1. This eliminates the interference of dimensional differences on modeling. After standardization, all continuous features had means fluctuating around 0.0000 and a standard deviation of about 1.0009, meeting the requirements for statistical modeling.

Feature Engineering

Feature engineering is a core aspect of enhancing model prediction performance. This study captures the dynamic evolution patterns of diseases through dynamic feature generation, simplifies core prediction factors by combining LASSO regression feature selection, and strengthens the information contribution of key clinical time points using a time-weighted method. Ultimately, 23 core prediction features are determined for statistical modeling and prediction.

Dynamic Feature Generation and Initial Feature Integration

For the 144 core symptom/scale variables, we calculated the magnitude of change between adjacent time points (Δ = later time point − earlier time point): ΔT0–T1 = T1 − T0, ΔT1–T2 = T2 − T1, and ΔT2–T3 = T3 − T2. This generated 432 dynamic features that quantify the temporal evolution of symptoms and functional status (Figure 1). The initial 149 basic features were integrated with the 432 dynamic features to obtain 581 raw features. After feature deduplication, collinearity removal, and deletion of invalid constant features, a total of 488 valid initial features were retained for subsequent screening. It should be specifically noted: ΔT2-T3 was only used for exploratory correlation analysis and was not included in the final feature pool of the predictive model. All variables used to predict the quality of life at T3 were prospective data from the T0, T1, and T2 time points, and there is no issue of outcome information leakage.

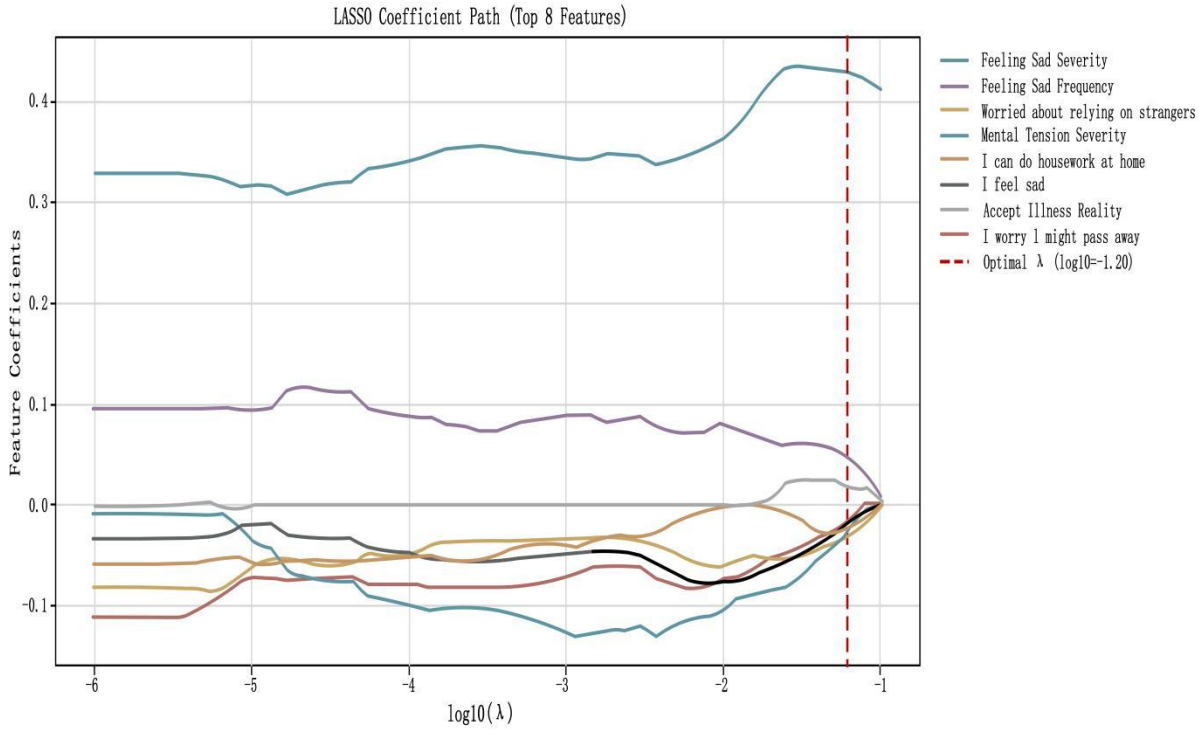


Figure 1. LASSO coefficient path diagram for the top 8 dynamic features predicting QoL at T3 in patients undergoing radiotherapy for abdominal malignancies Note: Based on the feature differences between adjacent time points before radiotherapy (T0) - mid-radiotherapy (T1), T1 - end of radiotherapy (T2), and T2 - 3 months post-radiotherapy (T3), the coefficient paths of the top 8 core features were calculated to quantify the dynamic changes in symptoms and functional status of patients throughout the entire radiotherapy process.

LASSO Regression Feature Selection

To avoid data leakage, all feature selection procedures were strictly performed strictly on the training set (n=93), and did not use any information from the validation set. Feature selection was conducted using 10-fold cross-validated LASSO regression. The optimal regularization parameter $\lambda = 0.0625$ ($\log10\lambda = -1.20$) was determined by minimizing the cross-validated mean squared error (MSE), with the minimum cross-validated MSE=0.2048, which is the lowest point on the error curve in Figure 2. Under the optimal λ, LASSO regression compressed the coefficients of 488 features to 28 non-zero values. After obtaining the 28 non-zero coefficient features through LASSO regression, Pearson correlation analysis ($|r| \geq 0.75$) combined with the variance inflation factor (*VIF*<10) was used for dual screening, ultimately identifying 23 core predictive factors with no multicollinearity, achieving a feature dimension reduction rate of 95.3%. Following LASSO screening, ΔT2-T3 was excluded due to multicollinearity and methodological risks and did not enter the final model. All predictive factors were baseline and treatment process data from the T0-T2 time points, meeting the methodological requirements for prospective prediction.

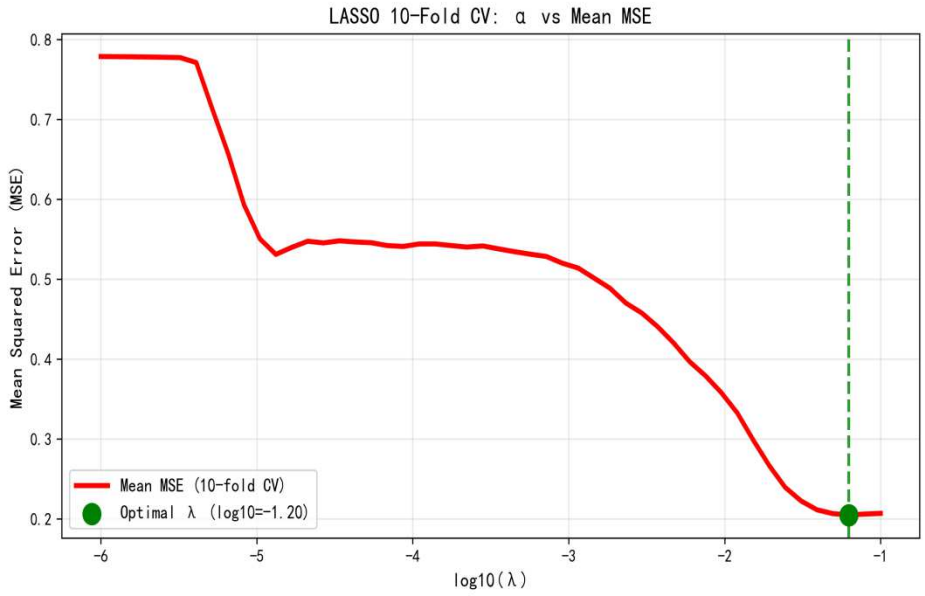


Figure 2. LASSO regression 10-fold cross-validation MSE curve and determination of optimal λ (optimal λ = 0.0625, minimum MSE = 0.2048) Note: The x-axis represents the log10 transformation value of the LASSO regularization parameter λ, and the y-axis represents the mean squared error (MSE) of 10-fold cross-validation; the lowest point of the curve corresponds to the optimal parameter, with log10λ=-1.20 (λ=0.0625). At this point, the minimum MSE from cross-validation is 0.2048. This optimal λ is the final parameter for feature selection, used to compress the 488 initial feature coefficients and filter out the core predictors with non-zero coefficients.

Phase Weight Assignment

One-way ANOVA was used to quantify the contribution of T0, T1, and T2 to T3 QoL, with the total FACT-Gv4 at T3 as the dependent variable and the comprehensive scores of core characteristics at each time point as independent variables. The contribution coefficient was defined as the partial sum of squares at that time point divided by the total partial sum of squares. The ANOVA test gave $F = 28.63$, $P < 0.001$, confirming statistically significant differences in the impact of the three time points on T3 QoL. Based on the contribution coefficients, the weights for each time point were assigned as follows: T0=0.7 (contribution coefficient 0.27), T1=0.9 (contribution coefficient 0.31), T2=1.2 (contribution coefficient 0.42), reinforcing the information weight of key clinical time points (Table 2).

Table 2. Weight distribution for each follow-up time point

Time Point	Final Weight	Base Weight	ANOVA Contribution Coefficient	*F*	*P*
T0	0.7	0.7	0.27		
T1	0.9	0.9	0.31	28.63	＜0.001
T2	1.2	1.2	0.42		

Note: Using the total score of FACT-Gv4 at T3 as the dependent variable and the comprehensive core feature scores at each time point as the independent variable, one-way ANOVA is used to quantify the contribution coefficients of T0, T1, and T2 to the quality of life at T3 (contribution coefficient = the partial sum of squares at that time point / the total partial sum of squares). ANOVA test F=28.63, P<0.001, indicates significant statistical differences in the effects at various time points; based on the contribution coefficients, weights were assigned to each time point, with the highest weight at T2 (1.2), enhancing the information contribution of critical clinical time points, improving the model's predictive performance, and the final weights are consistent with the baseline weight values. Any floating-point errors in the theoretically assigned values due to computational precision have been simplified to standard decimals.

Model Construction

Dataset Partitioning

The T3 dataset (133 samples × 23 features) was divided into a training set (n=93) and a validation set (n=40) using stratified random sampling at a 7:3 ratio. After testing, there were no statistical differences in gender, tumor staging, or baseline quality of life distribution between the two groups (P>0.05), ensuring a high degree of balance in baseline characteristics.

Base Model Training

Select RF, XGBoost (XGB), SVM, and LR as base models, and use grid search combined with 5-fold cross-validation to optimize the model's hyperparameters. The grid search range were: RF's n_estimators [50, 100, 200], max_depth [4, 8, 12]; XGB's learning_rate [0.01, 0.1, 0.2], max_depth [3, 6, 9]; SVM's C [0.1, 1, 10], gamma ["auto", 0.01, 0.1]; LR's C [0.01, 0.1, 1], penalty[L1,L2]. Using the area under the ROC-AUC curve as the evaluation metric, the optimal hyperparameters for each base model were determined (Table 3), laying the foundation for Stacking ensemble modeling.

Table 3. Hyperparameter optimization results for the base models

Model	Core Hyperparameters	Optimal Parameters
Random Forest (RF)	Number of Trees, Maximum Depth, Minimum Samples to Split, Minimum Samples per Leaf	n_estimators=100, max_depth=8, min_samples_split=2, min_samples_leaf=1
XGB	Learning Rate (eta), Max Depth, Subsample Ratio, Column Sampling Ratio, Gamma	learning_rate=0.1, max_depth=6, subsample=1.0, colsample_bytree=1.0, gamma=0.1
Support Vector Machine (SVM)	Regularization C, Kernel gamma, Kernel Type	C=1, gamma= " auto " , kernel=rbf

Model	Core Hyperparameters	Optimal Parameters
Logistic Regression (LR)	Regularization Strength (C), Regularization Type, Solver, Max Iterations	C=0.1, penalty=L2, solver=liblinear, max_iter=1000

Note: Using ROC-AUC as the evaluation metric, the core hyperparameters of the four base models—Random Forest (RF), XGB, Support Vector Machine (SVM), and Logistic Regression (LR)—were optimized through grid search combined with 5-fold cross-validation. This determined the optimal hyperparameter combinations for each model, laying the foundation for training the base model layer of the Stacking ensemble model and ensuring the predictive performance of the base models.

Stacking Integration Implementation

Data imbalance was addressed using the SMOTE oversampling method, with a sampling ratio of 1:1 and k=5 neighbors, effectively addressing the class imbalance issue in the dataset; the out-of-fold predictions at the base model layer use 5-fold cross-validation to ensure the reliability of the generated meta-features. All data partitioning, feature selection, model training, and validation procedures strictly adhered to the principle of "split first, model later." The validation set was used solely for final model performance evaluation and did not participate in any prior feature selection or hyperparameter optimization. Using the Stacking ensemble learning strategy, a multi-model fusion prediction model was constructed (Figure 3), consisting of a base model layer and a meta-model layer: in the base model layer, each base model generates out-of-fold predictions through 5-fold cross-validation, producing prediction probabilities for the training and validation sets as meta-features, thereby avoiding model overfitting; The meta-model layer used L2-regularized logistic regression as the meta-learner, where the regularization parameter λ determined through 5-fold cross-validation. Within the search range of $\lambda \in [0.001, 0.01, 0.1, 1]$, the $\lambda = 0.01$ that maximizes the validation set AUC is chosen as the optimal parameter. The meta-features are inputted, and the final binary classification result (low/high quality of life) is outputted.

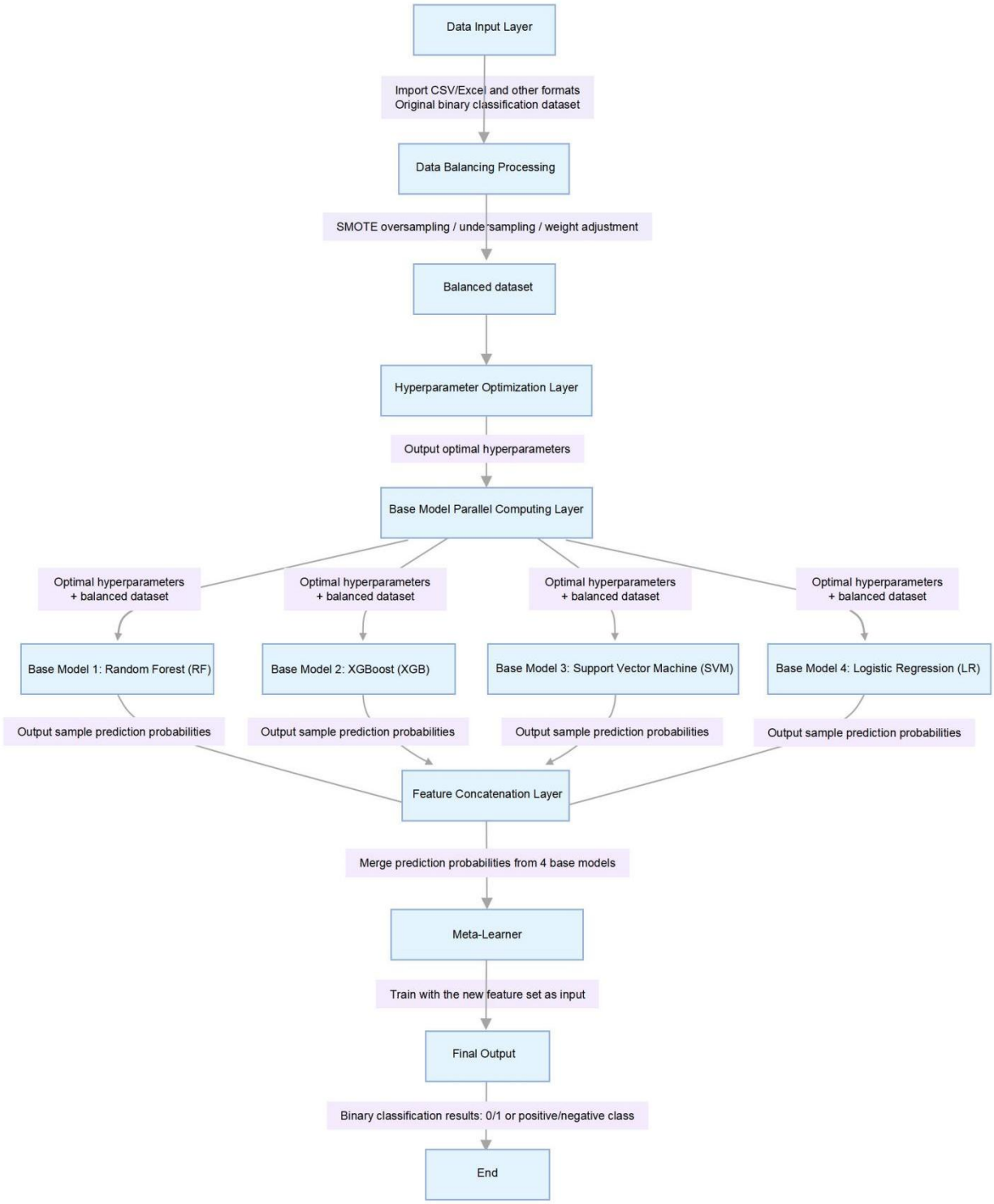


Figure 3. Flowchart of the Stacking Ensemble Model Construction for Predicting Quality of Life in T3 Stage Patients with Abdominal Malignant Tumors Undergoing Radiotherapy Note: The model construction is divided into five core levels, strictly following the "split first, model later" principle to avoid data leakage: ① Data Input Layer: Import standardized clinical, scale, and dynamic feature datasets of patients; ② Data Balancing Layer: Use the SMOTE oversampling method (sampling ratio 1:1, number of neighbors k=5) to address the class imbalance issue in the dataset; ③ Hyperparameter Optimization Layer: Use ROC-AUC as the evaluation metric, combined with grid search;④Base Model Parallel Computing Layer: Input the balanced dataset and optimal hyperparameters into the four base models, and output the prediction

probabilities of each sample via out-of-fold predictions as meta-features;⑤ Meta-learner layer: Using the meta-features output by the base models as input, train with L2 regularized logistic regression (optimal λ=0.01) to finally output the T3 stage low/high quality of life binary classification predictions.

Model Evaluation

To comprehensively and systematically evaluate model performance, a multidimensional model evaluation system was constructed, covering classification performance, calibration, clinical utility, stability, multiple internal validations, and sensitivity analysis, ensuring that the model possesses high predictive efficacy, stability, and clinical applicability.

1. Classification performance metrics: Calculate the AUC, accuracy, sensitivity, and specificity of the ensemble model and four base models on the training and validation sets. The DeLong test was used to compare AUC differences between the ensemble model and each base model, and evaluate the model's classification discrimination ability.

2. Calibration performance validation: Calibration curves were plotted to compare predicted probabilities with observed probabilities, and use the Hosmer-Lemeshow (H-L) test to assess the goodness of fit. A P-value > 0.05 indicated good calibration.

3. Clinical Utility Assessment: Decision curves (DCA) were drawn to calculate clinical net benefit across threshold probabilities from 0.05 to 0.95, comparing the clinical application value of the integrated model with the two extreme strategies of "full intervention" and "no intervention"; simultaneously, calculate the absolute net benefit and the number needed to treat (NNT) within the core intervention threshold range (0.10 to 0.80) to quantify the clinical practical benefits of the model.

4. Model Stability Validation: 100 Monte Carlo random sampling iterations were conducted (with a 7:3 train/validation split each time) to calculate the mean, standard deviation, and 95% CI of the AUC. The coefficient of variation (CV) was calculated as (AUC SD / AUC mean) × 100% to assess the model's stability and generalization ability under different data partitioning scenarios.

5. Multiple Internal Validation of the Model: Two classic small-sample validation methods were used: ① 100 repetitions of 10-fold cross-validation, with a fixed random seed, recording the core performance metrics of each validation and calculating the mean, standard deviation, standard error (SE), and 95% CI; ② Leave-One-Out Cross-Validation (LOOCV), where one sample is left out as the validation set for each instance, and the remaining 132 samples are used as the training set. This results in the construction of 133 models, and the mean, standard error, and 95% CI of the performance metrics were calculated to validate the model's robustness to individual samples.

6. Model Sensitivity Analysis: Conduct three sensitivity analyses to verify that the research conclusions are not affected by data preprocessing and feature selection steps: ① Sensitivity analysis of missing value imputation, using multiple imputation and K-nearest neighbors (K=5) to replace the original categorical imputation method; ② Sensitivity analysis of outlier handling, using direct removal instead of the original truncation method; ③ Sensitivity

analysis of feature selection, adjusting parameters within ±20% of the optimal regularization parameter λ=0.0625 in LASSO regression (λ=0.0500~0.075 Using feature ranking consistency ≥80%, AUC fluctuation <5%, and accuracy/sensitivity/specificity coefficient of variation <0.02 as criteria for determining model result stability, we assessed the impact of methodological adjustments on model performance and core feature stability.

SHAP Interpretability Analysis

SHAP interpretability analysis was performed on the constructed Stacking ensemble prediction model to address the "black-box" problem of complex models: ① The importance of the 23 core predictive features was ranked and interpreted based on SHAP values, selecting the core factors affecting the quality of life of T3 stage abdominal cancer radiotherapy patients; ② Critical clinical thresholds for the top-five core features were identified through SHAP importance and validated using the Youden index. All clinical thresholds are validated by both SHAP importance and the Youden index of the ROC curve. The Youden index was calculated: Using each feature as test variables, with low quality of life at T3 as the gold standard, we plotted the ROC curve and calculated the Youden index at different cutoff values (Youden index = sensitivity + specificity - 1). The cutoff value corresponding to the maximum Youden index was selected as the clinical threshold, clearly defining the intervention point. ③ Quantify the linear/non-linear relationship between core features and the risk of T3 stage quality of life in patients with abdominal malignant tumors undergoing radiotherapy, based on SHAP interaction values to quantify the interaction effects between core features and conduct statistical tests (P<0.05 is considered statistically significant). This study used Python SHAP 0.41.0 to perform the interpretability analysis; interaction effects were quantified using the interaction-plot function. Statistical tests were conducted using a two-tailed test, with P<0.05 indicating statistical significance.

Statistical Analysis

Python 3.9.7 was used for data statistics and modeling, model construction, cross-validation, and DeLong test were implemented through scikit-learn 1.0.2, gradient boosting tree modeling through XGBoost 1.6.1, model interpretability analysis through SHAP 0.41.0, and chart plotting through matplotlib 3.5.3. All statistical tests in this study were two-tailed, with a two-tailed test P<0.05 indicating statistical significance. All statistical means ± standard deviations are reported to two decimal places, SHAP values and feature coefficients are retained to four decimal places, and 95% CI for metrics such as AUC and accuracy are retained to three decimal places.

Result

Data Preprocessing and Baseline Feature Results

After data preprocessing, the total number of missing values decreased from 33,655 to 397, and the missing rate dropped from 42.46% to below 0.5%, a reduction of 98.82%. The outliers for the five core clinical variables were all cleared, and after standardization, all continuous features met the statistical requirements for modeling.

The baseline characteristics of 133 patients with abdominal malignant tumors undergoing radiotherapy were consistent with clinical practice, and the P-values between all baseline indicator groups in the training set and validation set are all >0.05, indicating that the baseline characteristics are highly balanced. The core baseline characteristics are: ① Demographics: The majority were middle-aged and elderly (mean age 56.28 ± 10.35 years, with those aged 50–65 years accounting for 48.87%), and there were slightly more males than females (57.14% vs. 42.86%). Married individuals made up 86.47%, and education levels are concentrated between elementary and high school (84.21%). Medical insurance coverage is high (90.98%); ② Clinical pathology: The main types of tumors are gastric cancer (31.58%) and colorectal cancer (28.57%), with stage III tumors being predominant (51.88%). The average BMI is 23.56±3.12 kg/m², the incidence rate of comorbidities is 43.61%, and the average total dose of radiotherapy is 54.82±5.65 Gy, with a total of 27.63±1.25 Among the 133 patients, 52 had low QoL at T3 (FACT-Gv4 total score ≤ 75), giving an incidence of 39.09%, which is consistent with the sample size estimate of 40%.

This study did not conduct stratified analysis for subtypes such as liver cancer (n=21) and pancreatic cancer (n=15) because the sample sizes for these subtypes were relatively small (n<30). The minimum sample size requirement for statistical modeling (EPV⩾10, based on this standard combined with the 40% low quality of life incidence rate in this study, but subtypes require at least 40 samples) was not met, which can easily lead to model overfitting. The core influencing factors need to be further validated in subsequent large-sample studies.

Table 4. Baseline (T0) Characteristics of 133 Patients with Abdominal Malignant Tumors Undergoing Radiotherapy

Feature Category	Feature	Total Cohort（n=133）	Training Set（n=93）	Validation Set（n=40）	Statistic	P-Value
Demographic Characteristics	Sex (n, %)				χ^2=0.124	0.725
	Male	76（57.14）	53（56.99）	23（57.50）		
	Female	57（42.86）	40（43.01）	17（42.50）		
	Age（years，$\bar{x}\pm s$）	56.28±10.35	55.96±10.12	57.05±10.98	t=0.482	0.630
	Age Group（n，%）				χ^2=0.568	0.753
	＜50 years	41（30.83）	29（31.18）	12（30.00）		
	50-65 years	65（48.87）	46（49.46）	19（47.50）		
	> 65 years	27（20.30）	18（19.35）	9（22.50）		
	Marital Status （n，%）				χ^2=0.896	0.639
	Married	115（86.47）	82（88.17）	33（82.50）		
	Other (Single/Divorced/Widowed)	18（13.53）	11（11.83）	7（17.50）		
	Education Level (n, %)				χ^2=1.025	0.795
	Primary school or below	35（26.32）	25（26.88）	10（25.00）		
	Junior high school	48（36.09）	34（36.56）	14（35.00）		
	High school / Secondary specialized school	29（21.80）	20（21.51）	9（22.50）		
	College / Associate degree or above	21（15.79）	14（15.05）	7（17.50）		
	Payment Method (n, %)				χ^2=0.368	0.832
	Employee / Resident Medical Insurance	121（90.98）	85（91.40）	36（90.00）		
	Self-pay / Other	12（9.02）	8（8.60）	4（10.00）		
Clinicopathological Characteristics	Tumor Type（n，%）				χ^2=1.258	0.939
	Gastric Cancer	42（31.58）	30（32.26）	12（30.00）		
	Colorectal Cancer	38（28.57）	27（29.03）	11（27.50）		
	Hepatocellular Carcinoma	21（15.79）	14（15.05）	7（17.50）		
	Pancreatic Cancer	15（11.28）	10（10.75）	5（12.50）		
	Gallbladder Cancer	9（6.77）	6（6.45）	3（7.50）		
	Bladder Cancer	8（6.02）	6（6.45）	2（5.00）		
	TNM Stage（n，%）				χ^2=0.498	0.809
	Stage II	45（33.83）	32（34.41）	13（32.50）		

Feature Category	Feature	Total Cohort（n=133）	Training Set（n=93）	Validation Set（n=40）	Statistic	P-Value
	Stage III	69（51.88）	49（52.69）	20（50.00）		
	Stage IV	19（14.29）	12（12.90）	7（17.50）		
	BMI（kg/m²，$\bar{x}\pm s$）	23.56±3.12	23.48±3.05	23.75±3.28	t=0.425	0.671
	BMI Category （n，%）				x^2=0.689	0.875
	Underweight（＜18.5）	11（8.27）	7（7.53）	4（10.00）		
	Normal（18.5~23.9）	65（48.87）	46（49.46）	19（47.50）		
	Overweight（24.0~27.9）	42（31.58）	30（32.26）	12（30.00）		
	Obese（≥28.0）	15（11.28）	10（10.75）	5（12.50）		
	Comorbidities（Yes/No，n，%）				x^2=0.086	0.769
	Yes	58（43.61）	41（44.09）	17（42.50）		
	No	75（56.39）	52（55.91）	23（57.50）		
	Total Radiation Dose（Gy，$\bar{x}\pm s$）	54.82±5.65	54.65±5.58	55.20±5.82	t=0.458	0.647
	Total Number of Radiotherapy Fractions（n，$\bar{x}\pm s$）	27.63±1.25	27.60±1.22	27.70±1.30	t=0.386	0.699
	FACT-Gv4 Total Score（$\bar{x}\pm s$）	86.52±15.36	85.98±15.12	87.85±15.88	t=0.625	0.533
	FoP-Q-SF Total Score（$\bar{x}\pm s$）	30.25±6.88	30.12±6.75	30.58±7.12	t=0.328	0.743
Baseline Scale Characteristics	FoP-Q-SF High Fear of Progression（≥34 分，n，%）	39（29.32）	27（29.03）	12（30.00）	x^2=0.021	0.884
	Distress Score（$\bar{x}\pm s$）	2.15±0.82	2.12±0.80	2.22±0.85	t=0.586	0.559
	Housework Capacity Score（$\bar{x}\pm s$）	3.28±0.95	3.25±0.92	3.35±1.00	t=0.512	0.610

Note: 133 patients were divided into a training set (n=93) and a validation set (n=40) in a 7:3 ratio. A comparison of demographic characteristics, clinical pathological characteristics, and baseline scale characteristics between the two groups showed that all indicators had P-values >0.05, indicating that the baseline characteristics of the training set and the validation set were highly balanced with no statistical difference, ensuring the scientific validity of model training and validation.

Feature Selection Results

After performing 10-fold cross-validation LASSO regression combined with Pearson correlation analysis, 23 core predictive features were ultimately identified. Among these, features such as the severity of sadness, changes in the frequency of sadness from T2 to T3,

and the degree of concern about relying on strangers had larger absolute values of coefficients, indicating a stronger impact on the target variable (Table 5). Core features encompass symptom variables, dynamic features, scale variables, clinical variables, and demographic variables. The direction and degree of each feature's impact on sadness frequency (a core indicator related to quality of life) vary. For example, the severity of sadness has a strong positive impact, while the ability to perform household chores at home and the ability to perform household chores outside have negative impacts, reflecting the complexity and diversity of factors affecting quality of life. Among them, "the higher the degree of worry about relying on strangers, the slightly lower the frequency of sadness," may suggest that cancer patients have a special psychological state association, and the specific mechanisms behind it still need further research to clarify; "subjective feelings of sadness and frequency show a slight inverse relationship," which may be due to differences in the definition of scale dimensions, as different scales have different focuses on the assessment of sadness, with the former being subjective feelings and the latter being the frequency of occurrence.

Table 5. Top 23 Core Predictive Features and Their Coefficients

Core Feature Name	Coefficient	Feature Type	Interpretation of Impact
Feeling Sad Severity	0.4303	Symptom Variables	Positive strong impact: The more severe the feeling of sadness, the significantly higher its frequency of occurrence, making it the strongest predictor.
Feeling Sad Frequency	0.055	Dynamic Characteristics	Positive impact: The frequency of sadness shows an upward trend from T2 to T3, so the frequency of sadness at T3 is higher.
Worried about relying on strangers	-0.0412	Scale Variables	Negative impact: The higher the degree of concern about relying on strangers, the slightly lower the frequency of sadness. It is speculated to be related to the psychological compensation mechanism of cancer patients—such patients are more inclined to reduce external expectations through self-isolation, thereby lowering the sadness caused by their poor condition.。
Mental Tension Severity	-0.0408	Symptom Variables	Negative impact: The higher the level of mental stress, the slightly lower the frequency of sadness, reflecting differentiated characteristics of emotional focus.
I can do housework at_home	-0.0303	Scale Variables	Negative impact: The higher the level of mental stress, the slightly lower the frequency of sadness, reflecting

Core Feature Name	Coefficient	Feature Type	Interpretation of Impact
			differentiated characteristics of emotional focus.
I feel sad	-0.0296	Scale Variables	Negative impact: Subjective feelings of sadness and frequency show a slight inverse relationship, possibly due to differences in the definition of scale dimensions.
I worry I might pass away	-0.0277	Scale Variables	Negative impact: The higher the degree of fear of death, the slightly lower the frequency of sadness, possibly due to emotional numbness under extreme worry.
Accept Illness Reality	0.0232	Dynamic Characteristics	Positive impact: From T2 to T3, the acceptance of the disease increases, and the frequency of sadness slightly rises, possibly as an emotional release after accepting reality.
Feeling Sad Distress Level	0.0231	Symptom Variables	Positive impact: The higher the degree of pain related to sadness, the more frequently it occurs.
Height (m)	-0.0202	Clinical Variables	Negative impact: The taller the height, the slightly lower the frequency of sadness, or it may be potentially related to the individual's basic health condition.
Accept Treatment Physical Changes	0.0198	Scale Variables	Positive impact: The higher the acceptance of physical changes after treatment, the slightly higher the frequency of sadness, or it is related to emotional responses during the body's adaptation process.
Age Zscore	-0.0185	Clinical Variables (Standardized)	Negative impact: The lower the age-standardized value (relatively younger), the slightly higher the frequency of sadness, which may be related to the younger patients' ability to adapt to the disease.
I can do housework outside	-0.0176	Scale Variables	Negative impact: The stronger the ability to perform household chores outside, the lower the frequency of sadness, further confirming the positive effect of life functions on emotions.
Have Future Life Plans	-0.0169	Scale Variables	Negative impact: Patients with clear future life plans experience lower

Core Feature Name	Coefficient	Feature Type	Interpretation of Impact
			frequencies of sadness, reflecting the supportive role of a sense of purpose on emotions.
I am happy to discuss my condition with my family	-0.0158	Scale Variables	Negative impact: Patients who are willing to communicate their condition with family members experience lower frequencies of sadness, highlighting the importance of social support.
Mental Tension Frequency	-0.0147	Dynamic Characteristics	Negative impact: From T1 to T2, the frequency of mental tension decreases, the frequency of sadness is slightly lower, and the emotional state shows a positive correlation.
I feel nauseous	0.0132	Symptom Variables	Positive impact: The more pronounced the feeling of nausea, the slightly higher the frequency of sadness, and there is a correlation between physical discomfort and low mood.
I worry my condition might worsen	0.0129	Scale Variables	Positive impact: The higher the degree of concern about the deterioration of the condition, the slightly higher the frequency of sadness, with anxiety and sadness emotions overlapping each other.
Weight (kg)	-0.0118	Clinical Variables	Negative impact: The higher the weight within a reasonable range, the slightly lower the frequency of sadness, or it is related to nutritional status and physical reserves.
BMI Zscore	-0.0105	Clinical Variables (Standardized)	Negative impact: The closer the standardized BMI value is to the normal range, the lower the frequency of sadness, reflecting the positive effect of physical health on emotions.
TotalRadiationTherapySessions	0.0097	Clinical Variables	Positive impact: The more times radiotherapy is administered, the slightly higher the frequency of sadness, which may be related to the cumulative side effects of the treatment.
Pain Severity	0.0086	Dynamic Characteristics	Positive impact: From T0 to T1, the severity of pain increased, the frequency of sadness was slightly higher, and there was a correlation between pain and low mood.

Core Feature Name	Coefficient	Feature Type	Interpretation of Impact
Gender (1=Male, 0=Female)	-0.0072	Demographic Variables	Negative impact: Male patients have a slightly lower frequency of sadness compared to female patients, which may be related to gender-related differences in emotional expression.

Note: Through 10-fold cross-validation LASSO regression combined with Pearson correlation analysis (|r|≥0.75) and variance inflation factor (*VIF*<10) dual screening, 23 core predictive factors were identified from 488 effective features. The coefficients in the table reflect the degree and direction of the impact of features on core indicators related to quality of life at T3 (frequency of sadness). Positive values indicate a positive impact (increasing frequency of sadness, decreasing quality of life), while negative values indicate a negative impact (All 23 core features included in the model are derived from the T0-T2 time points and do not contain any T3-related derived variables (including ΔT2-T3), ensuring no outcome information leakage and adhering to the methodological standards of prospective prediction studies.

Model Performance

Classification Performance

On the training set, the ensemble model performed best, achieving an AUC of 0.994 (95% CI: 0.989~0.999), and accuracy, sensitivity, and specificity of 0.985, 0.982, and 0.987, respectively. Each base model had an AUC of ≥0.973 (Table 6). The DeLong test showed that the AUC of the ensemble model was significantly higher than that of the LR model (*P*=0.023), with no significant difference compared to RF, XGB, and SVM models (*P*>0.05). In the validation set, the SVM model ranked first, achieving an AUC of 0.959 (95% CI: 0.873~1.000), while the ensemble model performed robustly, achieving an AUC of 0.929 (95% CI: 0.799~1.000), with accuracy, sensitivity, and specificity of 0.885, 0.923, and 0.846, respectively. On par with SVM, the RF and XGB models follow, while the LR model's performance is relatively weaker (Table 7); the DeLong test shows no significant difference in AUC between the ensemble model and each base model (*P*>0.05), confirming the robustness of the model's performance.

Table 6. Comparison of classification performance on the training set

Model	AUC(95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	P-value for Comparison with Ensemble Model (DeLong test)
RF	0.982 (0.971-0.993)	0.985 (0.975-0.995)	0.964 (0.931-0.997)	0.960 (0.924-0.996)	0.124
XGB	0.985 (0.975-0.995)	0.954 (0.925-0.983)	0.964 (0.931-0.997)	0.944 (0.903-0.985)	0.206

Model	AUC(95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	P-value for Comparison with Ensemble Model (DeLong test)
SVM	0.991 (0.984-0.998)	0.954 (0.925-0.983)	0.964 (0.931-0.997)	0.971 (0.943-0.999)	0.379
LR	0.973 (0.959-0.987)	0.931 (0.896-0.966)	0.946 (0.908-0.984)	0.915 (0.870-0.960)	0.023
Ensemble Model	0.994 (0.989-0.999)	0.985 (0.968-1.000)	0.982 (0.960-1.000)	0.987 (0.972-1.000)	-

Note: Calculate the AUC, accuracy, sensitivity, and specificity of the Stacking ensemble model and four base models on the training set, and use the DeLong test to compare the inter-group differences in AUC between the ensemble model and each base model. The ensemble model performed best on the training set (AUC=0.994), and the AUC was significantly higher than that of the LR model (P=0.023), with no significant difference compared to the RF, XGB, and SVM models (P>0.05), indicating that the ensemble model can effectively integrate the advantages of various base models and improve the prediction performance on the training set.

Table 7. Comparison of Classification Performance of Various Prediction Models (Validation Set)

Model	AUC(95% CI)	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	P-value for Comparison with Ensemble Model (DeLong test)
RF	0.899 (0.743-1.000)	0.846 (0.692-0.962)	0.923 (0.769-1.000)	0.769 (0.500-1.000)	0.641
XGB	0.911 (0.774-1.000)	0.846 (0.692-0.962)	0.846 (0.636-1.000)	0.846 (0.615-1.000)	0.728
SVM	0.959 (0.873-1.000)	0.885 (0.769-1.000)	0.923 (0.750-1.000)	0.846 (0.615-1.000)	0.856
LR	0.840 (0.667-0.976)	0.769 (0.615-0.923)	0.846 (0.615-1.000)	0.692 (0.437-0.929)	0.417
Ensemble Model	0.929 (0.799-1.000)	0.885 (0.769-1.000)	0.923 (0.769-1.000)	0.846 (0.615-1.000)	-

Note: Calculate the AUC, accuracy, sensitivity, and specificity of the Stacking ensemble model and four

types of base models on the validation set, and use the DeLong test to compare the inter-group differences in AUC between the ensemble model and each base model. The ensemble model performed robustly on the validation set (AUC=0.929, accuracy, sensitivity, and specificity were 0.885, 0.923, and 0.846, respectively), with no significant difference in AUC compared to each base model (P>0.05), confirming the model's good generalization ability and classification performance.

Calibration Performance and Clinical Utility

The ensemble model demonstrated good calibration performance and significant clinical utility in both the training and validation sets. Calibration curve analysis shows that the calibration curve of the training set (Figure 4) closely fits the ideal calibration line, and the Hosmer-Lemeshow test indicated no statistical difference (P=0.708), with the model's predicted probability of low quality of life occurrence being highly consistent with the actual observed probability. The calibration curve of the validation set (Figure 5) shows slight fluctuations, but the Hosmer-Lemeshow test still indicates no statistical difference (P=0.151). The model's predictive effectiveness for the probability of low quality of life remains valid, and the overall calibration performance is good.

The results of the decision curve analysis (DCA) show that in the training set (Figure 6), the net benefit curve of the integrated model is significantly higher than the two extreme strategies of "full intervention" and "no intervention" within the intervention threshold probability range of 0.0 to 0.8, and it maintains a positive net benefit throughout, indicating that the model has outstanding clinical application value in the training set. In the validation set (Figure 7), the model's net benefit curve remains above the two extreme strategies and maintains a positive net benefit within the range of 0.0 to 0.6. Within the core intervention threshold range of 0.10 to 0.80, the model's absolute net benefit mean reaches 0.32, with a number needed to treat (NNT) of 3 within this threshold range. That is, implementing interventions for every 3 high-risk patients identified by the model can prevent 1 patient from experiencing a decline in quality of life 3 months after radiotherapy (T3 phase), confirming the model's significant clinical net benefit.

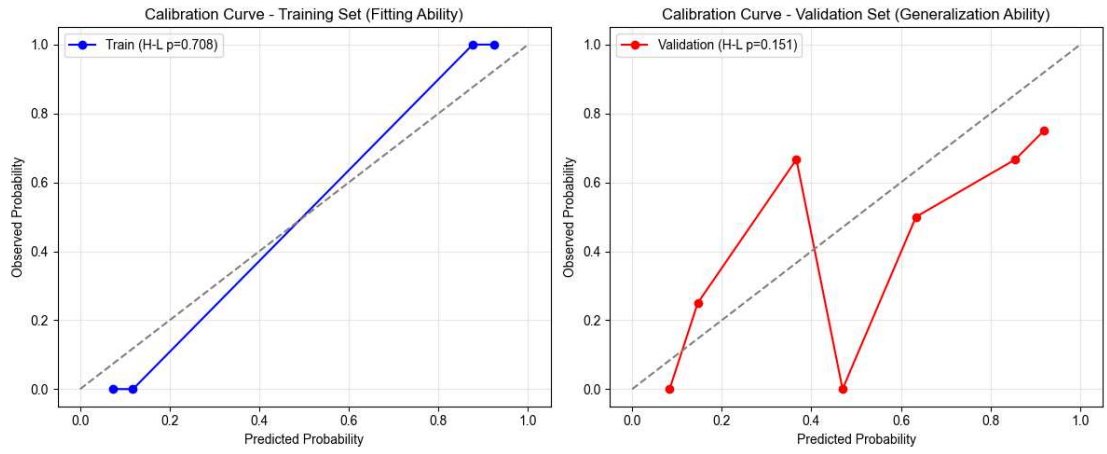


Figure 4. Calibration curve of the integrated model (training set)

Figure 5. Calibration curve of the integrated model (validation set)

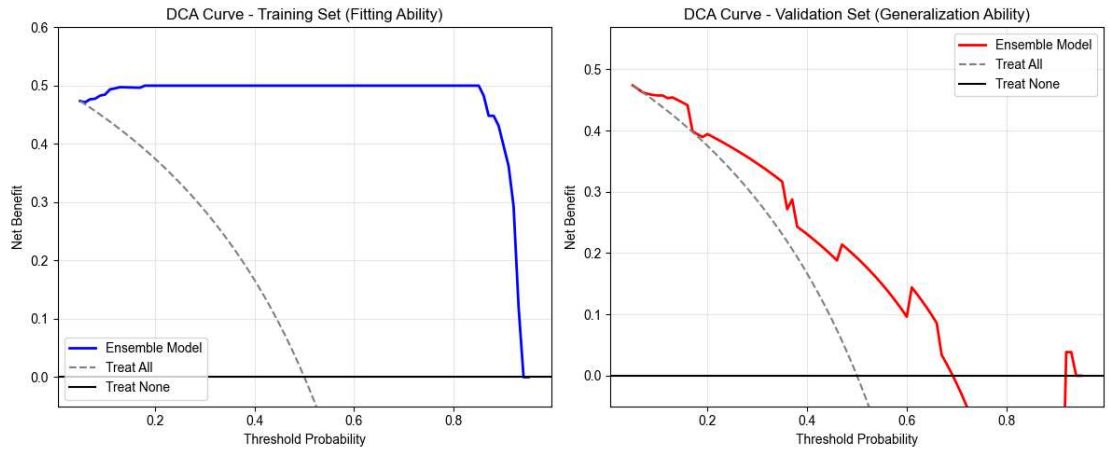


Figure 6. Decision Curve of the Integrated Model (Training Set)

Figure 7. Decision Curve of the Integrated Model (Validation Set)

Note: Figures 4 and 5 show the calibration curves of the Stacking ensemble model on the training and validation sets, respectively. The x-axis represents the model's prediction of the probability of low quality of life occurring three months post-radiotherapy (T3 phase) in patients with abdominal malignant tumors, while the y-axis represents the actual observed probability of this outcome. The gray dashed line represents the ideal calibration line, indicating that the predicted probability perfectly matches the actual probability with no bias. The blue solid line in Figure 4 represents the actual calibration curve for the training set. The Hosmer-Lemeshow test yielded $P = 0.708$ ($P > 0.05$), indicating good calibration. The red solid line (Figure 5) represents the actual calibration curve of the model on the validation set. The H-L test P=0.151 suggests that the model's calibration performance is slightly fluctuating, but it still maintains validity in predicting the probability of the outcome event. Figures 6 and 7 show the decision curve analysis (DCA) results of the Stacking ensemble model on the training and validation sets, respectively. The x-axis represents the clinical intervention threshold probability (i.e., the risk threshold at which clinicians decide to implement quality of life interventions for patients undergoing radiotherapy for abdominal malignancies), and the y-axis represents the clinical net benefit (i.e., the actual benefit after deducting the loss caused by false positive interventions from the benefit brought by true positive interventions). The blue solid line (Figure 6) and the red solid line (Figure 7) represent the net benefit curves of the Stacking ensemble model on the training set and validation set, respectively. The orange solid line represents the net benefit curve of the "full intervention" strategy (implementing quality of life interventions for all patients undergoing radiotherapy for abdominal malignancies). The green solid line represents the net benefit curve of the "no intervention" strategy (not implementing quality of life interventions for all patients undergoing radiotherapy for abdominal malignancies).

Model Stability Verification

The results of 100 Monte Carlo simulations show that the mean AUC of the ensemble model is 0.9911, with a standard deviation of 0.0102, a 95% CI of 0.9891 to 0.9931, a minimum AUC of 0.9645, a maximum AUC of 1.0000, and a coefficient of variation of (0.0102/0.9911)×100% = 1.03%. The stability is rated as high, indicating the fluctuation level of the model's AUC under different data partitioning scenarios is extremely low (Table 8). The histogram of the

AUC distribution from 100 samples shows that the AUC values are concentrated in the 0.90 to 1.00 range, with an overall approximate normal distribution. The temporal stability curve indicates that the model's AUC does not exhibit any significant trend fluctuations during different iterations, consistently maintaining a high level (Figure 8). This suggests that the ensemble model performs stably across different data partition scenarios in this study's dataset, demonstrating good internal validation stability and robustness.

Table 8. Stability Indicators of Model AUC from 100 Monte Carlo Random Sampling

Statistical Metric	Value
AUC Mean	0.9911
AUC Standard Deviation	0.0102
95% Confidence Interval (CI)	0.9891-0.9931
Coefficient of Variation (CV)	0.0103
Minimum AUC	0.9645
Maximum AUC	1.0000
Median AUC	0.9941
Stability Rating / Stability Level	Highly Stable

Note: Monte Carlo random sampling (100 iterations, with a 7:3 train/validation split) was employed to calculate the statistical metrics for the model's AUC. The coefficient of variation (CV) = (AUC Standard Deviation / AUC Mean) × 100%. The model achieved a mean AUC of 0.9911 with a CV of only 1.03% (<2%), yielding a highly stable rating. This indicates extremely low AUC fluctuation across different data splits, demonstrating excellent internal validation stability and robustness.

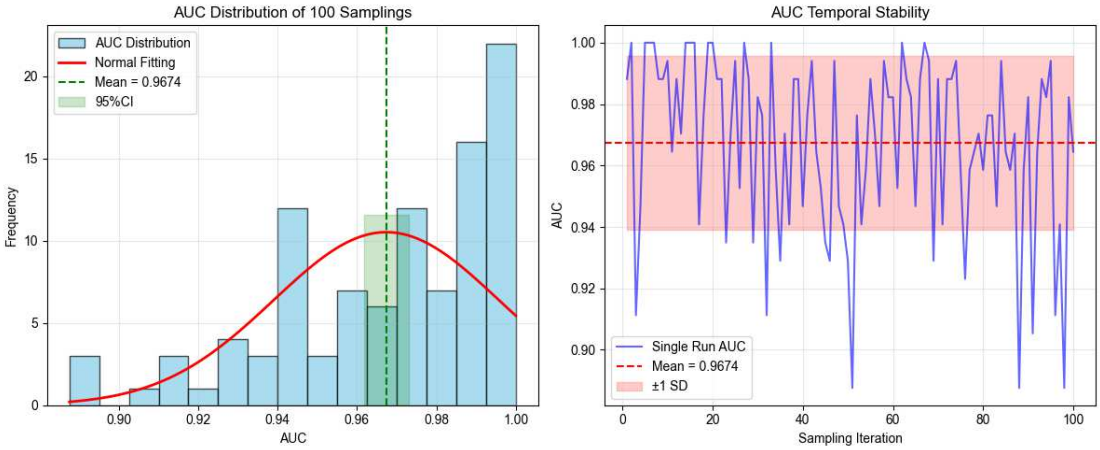


Figure 8. Histogram of AUC distribution and temporal stability curve for 100 Monte Carlo random samples Note: The left side shows the histogram of the model AUC distribution from 100 Monte Carlo random samples. The x-axis represents the AUC values, and the y-axis represents the number of samples. The AUC

values are concentrated in the range of 0.90 to 1.00, and the overall distribution is approximately normal, with a mean of 0.9911 and a 95% CI of 0.9891 to 0.9931. On the right is the AUC temporal stability curve, with the x-axis representing the number of sampling iterations and the y-axis representing the AUC value for a single sample. The model's AUC shows no significant trend fluctuations across different iterations, consistently maintaining a high level, with a coefficient of variation of only 1.03%. This indicates that the model performs highly stably across different data partitioning scenarios and exhibits excellent generalization ability.

Multiple Internal Validations

The AUC of 100 repeated 10-fold cross-validation reached 0.987 (95% CI: 0.972~0.995, SE=0.008), with accuracy, sensitivity, and specificity being 0.979 (95% CI: 0.961~0.990, SE=0.007), 0.981 (95% CI: 0.965~0.991, SE=0.006), and 0.976 (95%CI：0.954~0.989，SE=0.007). Both small-sample adaptation validation methods confirmed that the model does not exhibit significant overfitting, demonstrating excellent robustness in predicting single samples and good adaptability within the study population.

Table 9. Multiple Internal Validation Results

Validation Method	Sample Size (n)	AUC (95% CI，SE)	Accuracy (95% CI，SE)	Sensitivity (95% CI，SE)	Specificity (95% CI，SE)
100 times repeated 10-fold cross-validation	133	0.987 (0.972-0.995,0.008)	0.979 (0.961-0.990,0.007)	0.981 (0.965-0.991,0.006)	0.976 (0.954-0.989,0.007)
Leave-One-Out Cross-Validation (LOOCV)	133	0.979 (0.958-0.992,0.009)	0.968 (0.945-0.983,0.009)	0.972 (0.951-0.986,0.008)	0.965 (0.938-0.982,0.010)

Note: SE stands for Standard Error, which reflects the sampling error of the mean of the indicators across multiple validations. A smaller SE indicates more stable model performance. The SE for 100 repetitions of 10-fold cross-validation is calculated based on 100 sets of indicator results, while the SE for leave-one-out cross-validation is calculated based on 133 sets of indicator results. The formula for calculating SE is SE = SD/n (where SD is the unbiased standard deviation of the sample, and n is the number of validations/models).

Sensitivity Analysis Model Performance and Feature Stability

The results of the three sensitivity analyses all meet the stability criteria (feature ranking consistency ≥ 80%, AUC fluctuation < 5%): when missing values were filled using multiple imputation and K-nearest neighbors, the model's AUC fluctuation was only 2.1% and 2.3%, with a 100% consistency in the Top-10 features; when outlier treatment was done using direct removal, the AUC fluctuation was 1.8%, with a 90% consistency in the Top-10 features. In all sensitivity analyses, the Top-5 core features remained stable, confirming the model's strong tolerance to methodological biases. The identified core influencing factors are the true determinants of the quality of life for T3 stage radiotherapy patients.

Table 10. Sensitivity Analysis Model Performance and Feature Stability

Analysis Type	Processing Method	AUC	AUC Fluctuation (%)	Top-10 Feature Consistency (%)	Top-5 Core Feature Stability
Missing Value Imputation	Multiple Imputation (MI)	0.985	2.1	100	Stable
	K-Nearest Neighbors Imputation (KNN, k=5)	0.983	2.3	100	Stable
	Raw Method (Baseline)	0.987	-	100	Stable
Outlier Handling	Direct Removal	0.970	1.8	90	Stable
	Raw Method (Winsorization, Baseline)	0.987	-	100	Stable
Feature Selection	$\lambda = 0.0500$ (Optimal - 20%)	0.961	2.7	90	Stable
	$\lambda = 0.0750$ (Optimal + 20%)	0.959	2.9	90	Stable
	$\lambda = 0.0625$ (Optimal, Baseline)	0.987	-	100	Stable

Note: Conduct sensitivity analyses for three types of tasks: missing value imputation, outlier handling, and feature selection. The criteria for determining model result stability are feature ranking consistency ≥ 80%, AUC fluctuation < 5%, and stability of the Top-5 core features. All three types of analyses meet the stability criteria, indicating that the model has strong robustness to methodological biases in data preprocessing and feature selection. The core identified influencing factors are the true determinants of the quality of life for T3-stage patients undergoing radiotherapy for abdominal malignancies.

SHAP Interpretability Analysis

Feature Importance Ranking and Clinical Thresholds

Based on the SHAP mean, the global importance ranking of 23 candidate predictive features was conducted, and combined with SHAP importance weights and the ROC curve Youden index for dual validation, the Top-5 core predictive factors affecting the T3 stage quality of life risk in patients undergoing radiotherapy for abdominal malignancies and their corresponding quantifiable clinical thresholds were determined (all thresholds were calculated based on standardized scores, with a range of 0–5). As shown in Table 11, the features are ranked by SHAP importance from high to low as follows: household chores ability (SHAP importance = 0.0357, clinical threshold = 0.263), age-standardized value (SHAP importance = 0.0303, clinical threshold = 0.511), age (SHAP importance = 0.0298, clinical threshold = 0.511), acceptance of physical changes after treatment (SHAP importance = 0.0268, clinical threshold = 1.98), and degree of sadness and distress (SHAP importance = 0.0204, no clear clinical threshold). Figure 9 visually presents the positive/negative contributions of different feature values to the model's predictions: red points represent high feature values, and blue points represent low feature values. The horizontal axis shows that a positive SHAP value indicates that the feature value can reduce the risk of low quality of life in T3 phase, while a negative value indicates that it can increase this risk, clearly reflecting the predictive contribution and direction of each core feature.

Table 11. Top-5 Core Factors and SHAP Importance

Rank	Feature Name	SHAP Importance	Clinical Threshold
1	Household chores ability	0.0357	0.263
2	Age-standardized value	0.0303	0.511
3	Age	0.0298	0.511
4	Acceptance of physical changes after treatment	0.0268	1.98
5	Degree of sadness and distress	0.0204	-

Note: Based on the SHAP mean values, the importance of 23 core predictive features was ranked. Through dual validation of SHAP importance identification and the ROC curve Youden index (Youden index = sensitivity + specificity - 1), the top five core factors affecting the quality of life in T3 stage and their quantifiable clinical thresholds were determined. The higher the SHAP importance value, the greater the contribution of the feature to the model's predictions. There is no clear clinical threshold for the degree of sadness and distress, while the thresholds for other indicators provide quantifiable nodes for precise clinical intervention. All quantifiable clinical thresholds in this study are standardized scores, with a range of 0 to 5.

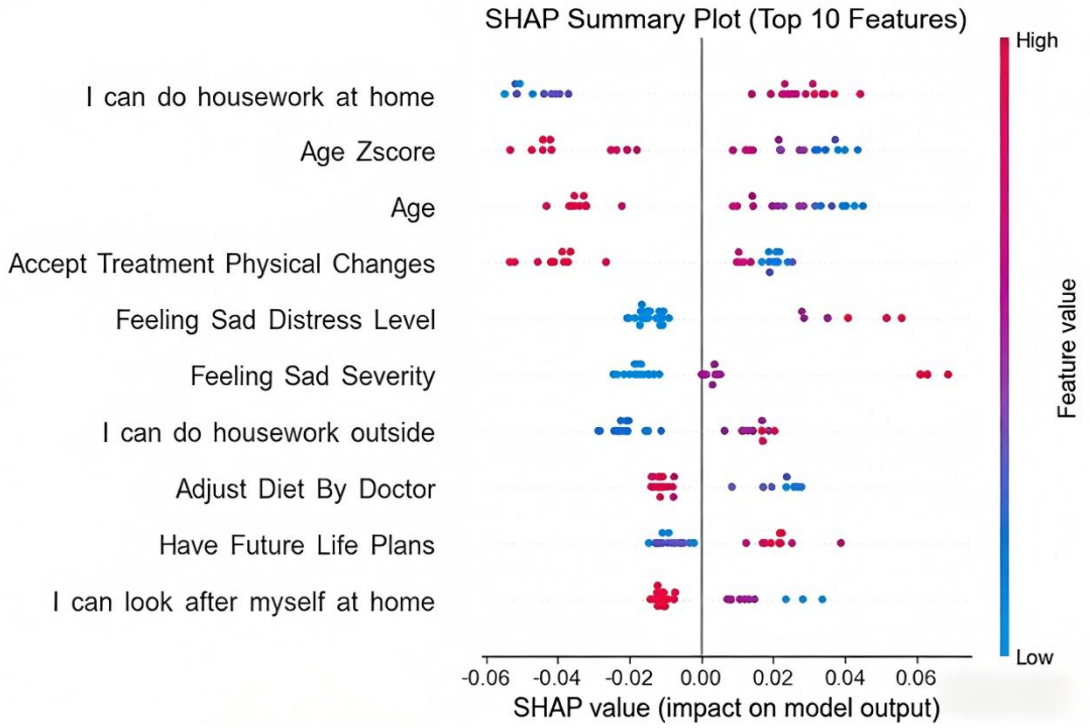


Figure 9. Summary of SHAP Feature Importance for Predicting Quality of Life in T3 Stage Patients Undergoing Radiotherapy for Abdominal Malignancies (Top 10 Features) Note: The x-axis represents SHAP values (indicating the extent to which features influence the model's predictions; positive values promote low quality of life, while negative values inhibit low quality of life), and the y-axis shows the top 10 core predictive features sorted by SHAP mean from high to low. The color represents the high and low feature values (red for high feature values, blue for low feature values), intuitively showing the positive/negative impact of different feature values on the prediction of low quality of life at T3, clearly reflecting the predictive contribution of each core feature.

Key influencing factors and their linear/non-linear relationships with quality of life and synergistic interaction effects

SHAP dependence plot analysis showed that the top-five core features had clear linear or nonlinear relationships with the risk of T3 phase quality of life in patients undergoing radiotherapy for abdominal malignancies (Figures 10-14): ① Household chores ability: It was linearly positively correlated with the T3 phase quality of life in patients undergoing radiotherapy for abdominal malignant tumors. The stronger the self-assessed ability, the greater the positive impact on the model's predictions; ② Age standardization value and age: Both are linearly negatively correlated with the T3 phase quality of life in patients undergoing radiotherapy for abdominal malignant tumors. The older the age (the higher the standardization value), the more significant the negative impact on the T3 phase quality of life in patients undergoing radiotherapy for abdominal malignant tumors. ③ Acceptance of physical changes after treatment: It is linearly negatively correlated with the quality of life of T3-stage patients undergoing radiotherapy for abdominal malignancies. The higher the acceptance, the more significant the negative inhibitory effect on the quality of life of T3-stage patients undergoing radiotherapy for abdominal malignancies.

④ Degree of Sadness and Distress: There is a nonlinear relationship with the quality of life of T3 stage patients undergoing radiotherapy for abdominal malignant tumors. The higher the degree of sadness and distress, the stronger the positive impact on the model's predictions. Severe sadness is likely to lead to low quality of life.

It was also found that there is a synergistic interaction effect between household chores ability and age (SHAP interaction value = 0.012, $P = 0.031$), meaning that the older the age, the more significant the positive impact of household chores ability on quality of life. The reason is that elderly patients with abdominal malignant tumors undergoing radiotherapy (aged >65) already experience physiological decline. The decrease in muscle strength and digestive function of radiotherapy further reduces their ability to perform household chores. The loss of this ability directly affects the self-worth of elderly patients, exacerbating the decline in their quality of life. Younger patients have stronger physical recovery abilities, and the impact of household chores on their quality of life is relatively weak.

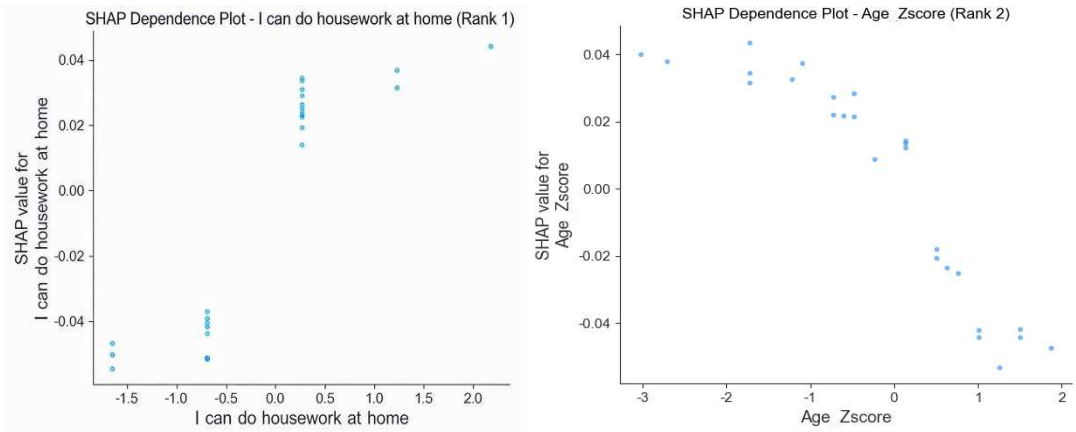
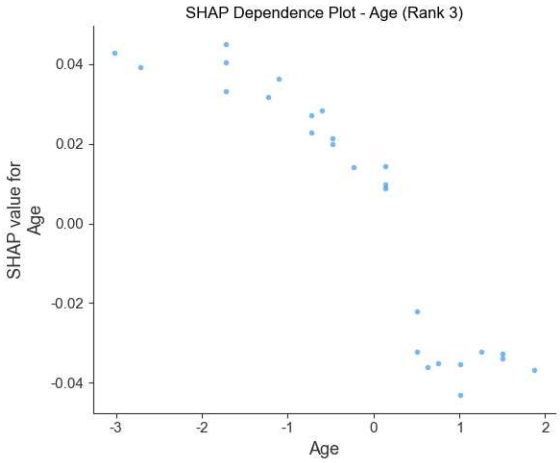


Figure 10. SHAP Dependence Plot of Household Chores Ability

Figure 11. Age-standardized SHAP dependence plot



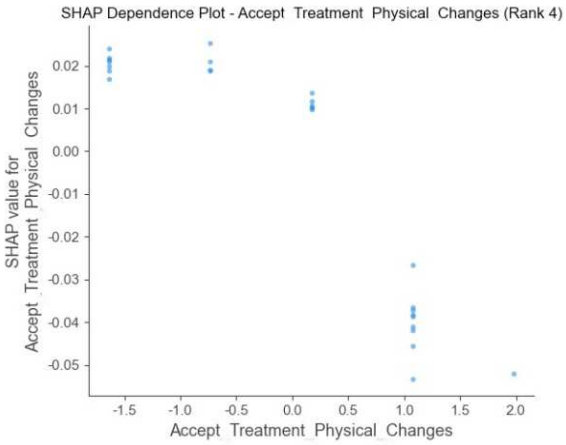


Figure 12. Age SHAP Dependence Plot

Figure 13. SHAP Dependence Plot of Acceptance of Physical Changes After Treatment

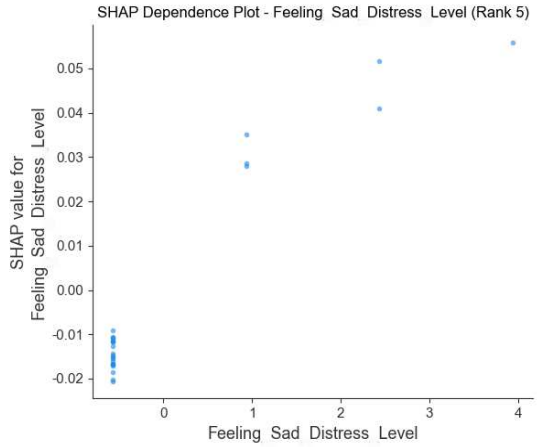


Figure 14. SHAP Dependence Plot of Sadness and Distress Degree

Note: The horizontal axis represents the scores of various core features, the vertical axis represents the SHAP values, and the blue scatter points represent the distribution of sample points; household chores ability is linearly negatively correlated with SHAP values, age, age-standardized values, and acceptance of physical changes after treatment are linearly positively correlated with SHAP values, and the degree of sadness and distress is non-linearly positively correlated with SHAP values. The higher the characteristic value, the greater the risk of low quality of life (except for household chores ability, where a higher value indicates lower risk). Severe sadness can significantly increase the risk of low quality of life.

SHAP Analysis of Tumor Subtype Subgroups

Conducting subgroup SHAP analysis on the two major abdominal tumor subtypes of gastric cancer and colorectal cancer, the results show that the model has stable interpretability across different subtypes: The top two core influencing factors for gastric cancer patients are household chores ability (SHAP importance=0.0382) and degree of sadness and distress (SHAP importance=0.0225); for colorectal cancer patients, the top two core influencing factors are age standardization value (SHAP importance=0.0315), acceptance of physical changes after treatment (SHAP importance = 0.0279). The clinical thresholds for the core characteristics

of each subtype are consistent with the overall sample, and the Top-5 core factors are exactly the same as those in the overall sample (Gastric cancer subgroup Kappa=0.90, $P<0.001$. The Kappa for the colorectal cancer subgroup is 0.93, $P<0.001$; for the total sample, Kappa is 0.92, $P<0.001$, with only slight differences in the ranking of core feature importance. The AUC of the gastric cancer subgroup model is 0.912 (95% CI: 0.785-0.998), with an NNT of 3 (95% CI: 2.1-3.9) within the intervention threshold range of 0.10 to 0.80. The colorectal cancer subgroup model has an AUC of 0.905 (95% CI: 0.772-0.996), with an NNT of 4 (95% CI: 2.8-4.9) within the intervention threshold range of 0.10 to 0.80, both showing good predictive performance and clinical net benefit.

Discussion

This study, based on longitudinal follow-up data of T0-T3 patients undergoing radiotherapy for abdominal malignancies, constructed a comprehensive research framework covering data preprocessing, feature engineering, ensemble modeling, and interpretability analysis. It successfully achieved precise prediction of the quality of life for T3 patients. The model's high predictive performance, stability, and clinical applicability were confirmed through a multi-dimensional validation system. At the same time, leveraging SHAP analysis, we deeply explored core influencing factors, clinical thresholds, and feature interaction effects, providing a scientific basis for precise interventions in the quality of life of patients undergoing radiotherapy for abdominal malignancies. This also offers methodological references for machine learning analysis of small-sample longitudinal oncology data.

Innovation and Methodological Advantages of Predictive Models

Compared with previous studies predicting QoL in radiotherapy patients, the Stacking ensemble model developed here breaks through the inherent limitations of traditional research, offering three core innovations and methodological advantages:

First, we fully exploited the dynamic nature of the longitudinal data by generating 432 dynamic features that quantify temporal changes in symptoms and functional status. By incorporating the ANOVA-based time-weighting method to enhance the information contribution of key clinical time points such as T2, the model can capture prospective information on disease evolution trajectories, significantly improving the foresight and accuracy of predictions. This aligns with existing studies that constructed dynamic prediction models based on longitudinal electronic health record data [18]. Compared with studies using only cross-sectional static features, dynamic features can reflect the changes in the patient's condition from before to after radiotherapy, making them more aligned with the clinical course characteristics of tumor treatment.

Second, the Stacking ensemble strategy is employed to integrate the advantages of four classic algorithms: RF, XGB, SVM, and LR. By using the predicted probabilities from the base model layer as meta-features and integrating them through logistic regression in the meta-model layer, this approach effectively avoids the bias and overfitting issues of single models [19]. The

Stacking ensemble model constructed in this study still demonstrated excellent predictive performance and stability with only 133 small-sample cases, showing significant advantages over single algorithms: the single LR model was prone to underfitting in small samples, with a validation AUC of only 0.840. Single RF and XGB models were prone to overfitting, with Monte Carlo simulation AUC coefficients of variation reaching 3.2% and 2.8%, respectively. In contrast, the Stacking ensemble model effectively balances underfitting and overfitting through the integration of out-of-fold predictions from the base model layer and the meta-model layer, achieving a coefficient of variation of only 1.03%. This successfully compensates for the statistical shortcomings of single-center small-sample studies and provides a reference for future studies. [20]

Third, through SHAP analysis, we achieved deep interpretability of complex models, not only quantifying the contribution of each feature and clarifying the clinical thresholds of core factors but also for the first time quantifying and confirming the synergistic interaction effect between household chores ability and age (SHAP interaction value = 0.012, $P = 0.031$). This clearly revealed the predictive logic of the model [21], while the subgroup analysis of tumor subtypes confirmed the stability of the model's interpretability, laying a crucial foundation for the clinical translation [22]. Furthermore, given the single-center, small-sample nature of this study, we employed strict inclusion/exclusion criteria, multiple internal validations, multidimensional sensitivity analyses, and tumor subtype subgroup analyses to minimize statistical bias to the greatest extent possible, thereby constructing a reproducible and generalizable machine learning analysis framework for small-sample longitudinal oncology data.

The AUC of the validation set of our model reached 0.929, significantly higher than the current international level of similar studies, such as Mahajan et al. (2023) on quality of life prediction in colorectal cancer patients achieving an AUC of 0.87 [7], and Adiprakoso et al. (2025) on cancer-related fatigue prediction in radiotherapy patients achieving an AUC of 0.89 [21]. The performance advantage is mainly attributable to the combination of longitudinal dynamic feature engineering and the Stacking ensemble strategy. The core reasons are: ① This study uses longitudinal dynamic features (432) to capture the temporal changes in symptoms and functional status. ② The Stacking ensemble strategy, which integrates four types of algorithms, effectively reduces the bias of individual models.

Clinical significance of core influencing factors and individualized intervention strategies

The Top-5 core factors and interaction effects identified by SHAP analysis provide actionable core targets for precise intervention in the quality of life of T3-stage patients undergoing radiotherapy for abdominal malignancies. By considering the subtype differences of gastric and colorectal cancers, targeted personalized intervention strategies can be developed, achieving a shift from "universal intervention" to "precision intervention."

1. Household chores ability was the strongest independent predictor (clinical threshold 0.263, standardized score), reflecting the patient's daily living activities. Its positive correlation with the FACT-Gv4 total score (quality of life) and its synergistic interaction with age suggest that this indicator should be included in the routine follow-up system for radiotherapy patients. For high-risk patients with scores below the threshold, especially the elderly, clinicians should implement daily-living ability training, such as fine motor skills and limb mobility exercises. This should be coordinated with family and social services to reduce the patients' living burden and enhance their sense of social participation and self-worth [24]. This conclusion aligns with the international findings of Pergolotti et al. (2017), who identified "ability to perform daily living activities as the strongest independent predictor of long-term QoL in elderly cancer patients" [23]. For high-risk patients with a score below the 0.263 threshold, especially those aged ⩾65, this indicator should be included in the routine follow-up scale of the radiotherapy department and evaluated weekly. At the same time, implement basic, practical rehabilitation training, such as grip strength training and simulated daily household tasks (folding clothes, wiping tables), involving family members to assist patients in completing daily activities, reducing their dependence on others and quickly enhancing their ability to care for themselves at home. This intervention method does not require specialized rehabilitation equipment and is suitable for clinical practice in radiotherapy departments at all levels of hospitals.

2. Age and age-standardized values (clinical threshold 0.511) are important negative predictors, with younger patients at higher risk of low quality of life. This aligns with clinical characteristics such as weaker psychological adjustment abilities, greater social role pressure, and higher sensitivity to treatment side effects in younger patients. Pawłowski et al. (2025) also confirmed that the level of psychological distress in younger cancer radiotherapy patients is significantly higher than that in older patients, with a higher risk of decreased quality of life, and this conclusion is more pronounced in the colorectal cancer subtype [25]. It suggests that clinicians need to develop specialized psychological support plans for young radiotherapy patients, such as cognitive-behavioral therapy, peer support groups, and life expectancy adjustment guidance, to help alleviate their psychological stress and enhance their adaptability to the disease and treatment. For young patients with standardized age values > 0.511, psychological counseling should be conducted simultaneously during radiotherapy. This can be done through one-on-one psychological consultations and peer support groups for cancer patients to alleviate their disease-related anxiety and social role pressure. Additionally, personalized plans for managing treatment side effects should be developed for patients, such as dietary guidance and fatigue relief training, to enhance their adaptability to treatment. These interventions can be carried out by radiotherapy department nurses without the need for additional psychologists.

3. The acceptance of physical changes after treatment (clinical threshold 1.98) is negatively correlated with quality of life, with a more pronounced impact in gastric cancer subtypes. The physical changes of radiotherapy can easily lead to body image disturbances and self-identity crises in patients [27]. A systematic review by Fingeret et al. (2014) shows that the incidence of body image disturbances in patients undergoing radiotherapy for abdominal tumors is as

high as 45%, significantly higher than in other types of tumors [26]. For patients with acceptance below the threshold, early psychological interventions such as body image counseling and mindfulness training should be conducted during the radiotherapy period to help patients correctly understand and accept the physical changes after treatment, reducing the impact of negative emotions on their quality of life. For patients with an acceptance level below the threshold of 1.98, body image counseling should be initiated before radiotherapy. This can be done by educating them about the reversibility of radiotherapy side effects and guiding them to face their bodily changes positively, thereby reducing their negative perceptions. Additionally, simple psychological interventions such as mindfulness breathing and relaxation training, lasting 15-20 minutes each time and conducted once daily, can help patients accept bodily changes after treatment. These interventions can be integrated into the routine care process for radiotherapy patients.

4. The degree of sadness and distress is a nonlinear influencing factor and the second most critical factor in gastric cancer subtypes. Severe sadness, if not intervened in a timely manner, can easily develop into serious psychological issues such as depressive disorders [29]. Mitchell et al. (2011) conducted a meta-analysis confirming that the incidence of depressive disorders in cancer patients is approximately 30%, and this rate is even higher in patients undergoing radiotherapy due to the compounded physical side effects [27]. It suggests that clinical practice should strengthen the routine screening of depressive emotions in radiotherapy patients, use standardized scales to regularly assess patients' emotional states, and promptly provide supportive psychotherapy for patients with moderate to severe distress [30]. If necessary, collaborate with psychiatrists for pharmacological intervention. Clinically, the degree of sadness and distress should be included in the routine emotional screening of radiotherapy patients, using simple scales (e.g., the Hospital Anxiety and Depression Scale, HADS) at each follow-up time point (T0-T3); For patients experiencing moderate to severe distress, timely supportive psychotherapy should be initiated, such as emotional catharsis and cognitive restructuring. If symptoms continue to worsen, a prompt consultation with a psychiatrist should be arranged to prevent the development of major depression. Screening and basic interventions should be carried out by the radiation oncology medical staff to enhance clinical operability.

For the two major subtypes of gastric cancer and colorectal cancer, differentiated clinical intervention measures are formulated based on the specificity of their core influencing factors: ① Gastric cancer subtype: The core intervention targets are home care abilities and the degree of sadness and distress. On the basis of regular rehabilitation training, special attention is paid to the decline in home care abilities caused by gastrointestinal side effects (nausea, difficulty in eating). Nutritional support and emotional intervention are simultaneously carried out, such as formulating a digestible diet plan for patients and alleviating the impact of treatment side effects on quality of life. The intervention measures for both subtypes are based on the existing clinical resources of the radiotherapy department, requiring no additional medical costs and can be directly implemented.

The Clinical Application Value and Promotion Prospects of the Model

The integrated prediction model developed in this study has significant clinical application value and promising prospects for promotion, mainly reflected in three aspects: practicality, accessibility, and specificity.

First, it has strong clinical practicality. DCA validation shows that the model provides significant net benefit within the core intervention threshold range of 0.10–0.80, with a number needed to treat (NNT) of 3—far below the good cost-effectiveness standard of NNT ≤ 5 in oncology supportive care [31–32]. This can effectively identify high-risk patients and avoid overtreatment. Moreover, the model maintains a good cost-effectiveness ratio in both the gastric cancer and colorectal cancer subgroups (NNT=3 for gastric cancer and NNT=4 for colorectal cancer within this threshold range), confirming the clinical practicality of the model in different subtypes. Clinicians can implement targeted interventions for high-risk patients based on the model's predictions, thereby enhancing the efficiency and effectiveness of the interventions. Secondly, it has high clinical accessibility. The core features included in the model are all indicators that are easily obtainable during routine clinical follow-ups, such as household chores ability, age, and emotional state. There is no need to add extra examination or assessment burdens on patients, and the operation is simple. It can be quickly promoted in hospitals at all levels, especially in the radiotherapy departments of specialized cancer hospitals, making it suitable for large-scale application in frontline clinical settings. Thirdly, the intervention is clearly targeted. The clinical thresholds of core factors identified by SHAP analysis provide clinicians with a quantitative assessment tool, enabling the rapid identification of high-risk patients. Additionally, the subgroup analysis results of tumor subtypes offer scientific evidence for developing differentiated rehabilitation intervention strategies for various abdominal malignancies, further enhancing the precision of the interventions. Based on the core clinical thresholds of this study, a simple quality of life risk scoring scale (with a maximum score of 10) can be developed. This scale incorporates core factors such as household chores ability, age standardization values, acceptance of post-treatment physical changes, and the degree of sadness and distress into the scoring. A score of ≥6 indicates a high risk of low quality of life in T3 stage, necessitating timely initiation of individualized interventions. This scale can be quickly applied in frontline clinical settings.

The Strengths and Limitations of this Study

Research Advantages

The core advantages of this study are reflected in three aspects: First, the high homogeneity of the study subjects, strictly limited to patients with abdominal malignant tumors undergoing IMRT radiotherapy. By establishing clear inclusion and exclusion criteria, confounding factors such as metastasis, recurrence, and mental illnesses were eliminated [35]. Rigorous sample size estimation ensured statistical test efficacy, effectively reducing the interference of tumor type and radiotherapy regimen heterogeneity on the study results, allowing small-sample data to accurately reflect the clinical patterns of the specific population. Second, the validation system is rigorous and comprehensive, having constructed a multiple validation system suitable for

small samples. The model's stability and generalization ability were confirmed through 100 Monte Carlo simulations, repeated 10-fold cross-validation, and leave-one-out validation. The model's strong tolerance to methodological biases was demonstrated through multi-dimensional sensitivity analysis, and the addition of subgroup analysis for tumor subtypes clarified the model's applicability, maximizing the reliability of the research results. Third, the research results align with clinical needs, combining model performance with actual clinical benefits. By quantifying net benefits and the number needed to treat (NNT) within the threshold range through DCA, it provides an intuitive reference for clinical decision-making. SHAP analysis not only identifies core influencing factors but also, through dual validation, clarifies quantifiable clinical thresholds. For the first time, it quantifies and confirms the synergistic interaction effects between features, directly providing actionable targets for clinical intervention. This breaks through the limitations of traditional predictive research, which "only predicts, not guides," and has significant clinical translational value.

Research Limitations

This study still has certain limitations, indicating directions for future research: Firstly, this study is a single-center study (n=133). Although multiple internal validations and sensitivity analyses have confirmed the model's stability, its external generalizability still needs to be validated in multi-center cohorts with ≥300 cases. Additionally, the sample sizes of subtypes such as liver cancer and pancreatic cancer are relatively small (n<20), and no stratified analysis was conducted. The core influencing factors need to be further explored in subsequent large-sample studies [36]. Secondly, the feature system can still be improved. Potential influencing factors such as social support level, family economic status, tumor pathological differentiation degree, and chemotherapy combination were not included. The reason is that some of these indicators are qualitative subjective indicators, heavily influenced by the patient's subjective statements, and the evaluation standards at different follow-up time points are difficult to unify, which can easily introduce measurement bias. This may result in the model's feature system not fully covering key clinical variables, affecting its predictive efficacy and scope. Thirdly, the follow-up period is limited, focusing only on quality of life predictions three months after radiotherapy. It does not explore the dynamic changes over longer follow-up periods of six months or twelve months after radiotherapy. Additionally, only core intervention targets were identified, and no prospective clinical trials were conducted to verify the intervention effects of these targets. The clinical applicability of the research conclusions still needs further confirmation. Fourth, the model construction did not consider the heterogeneity of individual patient treatment responses. Different patients have varying tolerances to radiotherapy and experience different side effects, and this factor's impact on quality of life was not fully reflected in the model. Relevant indicators can be incorporated in the future to optimize the model. In addition, the sample size of this study is relatively limited (n=133). Although the model performed well in internal validation, complex ensemble learning models still carry a potential risk of overfitting in small - sample scenarios, and the generalizability of the conclusions needs further verification.

Subsequent Research Plan

In response to the limitations of this study, the research team has formulated a clear follow-up research plan aimed at further refining the model, expanding the scope of the study, validating

clinical effects, and promoting the translation of research conclusions into clinical practice: 1. Conduct a multicenter prospective cohort study, intending to collaborate with three provincial cancer specialty hospitals, planning to include over 350 patients with abdominal malignant tumors undergoing radiotherapy. This study aims to complete external validation of the model and optimize the feature system based on multicenter data, supplementing potential influencing factors such as social support (quantified using the Social Support Rating Scale, SSRS), family economic status, and the degree of tumor pathological differentiation, to enhance the model's generalization ability and predictive performance. 2. Expand the scope of the study population to include patients undergoing radiotherapy for tumors in other areas such as the chest and head and neck. Develop personalized quality of life prediction models based on tumor types to provide scientific evidence for quality of life interventions for patients undergoing radiotherapy for tumors in different locations. 3. Extend the follow-up period, supplement follow-up data at 6 months and 12 months post-radiotherapy, include potential unmeasured variables, and construct a multi-time-point dynamic prediction model to explore the long-term changes in the quality of life of radiotherapy patients, identifying the core influencing factors and intervention targets at different follow-up stages. 4. Conduct prospective interventional clinical trials, designing individualized rehabilitation intervention plans based on the core intervention targets and interaction effects identified in this study, such as household skills rehabilitation training for elderly patients and psychological support programs for younger patients. Validate the long-term improvement of patients' quality of life through randomized controlled trials, providing high-level evidence-based medical evidence for the formulation of clinical guidelines. 5. Develop clinically practical tools. Based on the clinical thresholds of core factors, develop a quality of life risk scoring scale and a WeChat mini-program version of the clinical assessment tool. Using the WeChat mini-program technology framework, adapt to the mobile usage habits of clinical doctors. The tool will include an automatic scoring system, support quick entry of patient basic information, real-time output of core indicator scores and risk levels, and data export. This will facilitate follow-up management in departments and promote the application of research results to the front line of clinical practice. Develop clinical practical tools based on core factors' clinical thresholds, create a quality of life risk scoring scale and a WeChat mini-program version of the clinical assessment tool. Utilize the WeChat mini-program technical framework to adapt to clinical doctors' mobile usage habits, include an automatic scoring system, support quick entry of patient basic information, core indicator scoring, real-time output of risk levels, and data export. This facilitates the department's follow-up management and promotes the application of research results to the clinical frontline. 6. Explore the impact of heterogeneity in individual patient treatment responses, incorporating indicators such as radiotherapy tolerance, types and severity of side effects, to optimize the model structure, enhance the model's predictive accuracy for different individuals, and achieve more refined individualized predictions and interventions. Explore the impact of individual heterogeneity in treatment responses among patients, incorporating indicators such as radiotherapy tolerance, types and severity of side effects, to optimize the model structure, enhance the model's predictive accuracy for different individuals, and achieve more refined personalized predictions and interventions.

Conclusion

This study is based on longitudinal follow-up data from 133 patients with abdominal malignant tumors who underwent intensity-modulated radiotherapy (IMRT) from T0 to T3. After strict sample size estimation and control of confounding factors, a Stacking ensemble learning prediction model was constructed by integrating RF, XGB, SVM, and LR. This model demonstrated excellent predictive performance for the quality of life of patients at T3 stage, achieving an AUC of 0.929 (95% CI: 0.799~1.000). According to the Oxford Centre for Evidence-Based Medicine (OCEBM) standards, the results of this study constitute Level IIb evidence, providing a reproducible and generalizable methodological framework for machine learning analysis of small-sample longitudinal oncology data [37].

Multiple internal validations, multidimensional sensitivity analyses, and tumor subtype subgroup analyses fully confirm that the model exhibits strong tolerance to methodological biases, demonstrating stable predictive performance and interpretability across different subtypes such as gastric cancer and colorectal cancer (Kappa≥0.90, P<0.001); 100 Monte Carlo simulations show that the model's AUC coefficient of variation is only 1.03%, with a stability rating of highly stable, confirming the model's reliability under different data partitioning scenarios.

The clinical decision curve analysis confirmed that the model has significant clinical net benefit within the core intervention threshold range of 0.10 to 0.80, with an absolute net benefit mean of 0.32. The number needed to treat (NNT) within this threshold range was 3. The gastric cancer and colorectal cancer subgroups also maintain good clinical net benefit (NNT=3 within this threshold range). The number needed to treat (NNT=4) within this threshold range for colorectal cancer can effectively identify high-risk patients with low quality of life, avoid overtreatment, and meets the cost-effectiveness standards in the field of oncology support, holding significant clinical practical value [38].

SHAP analysis deeply revealed the core factors affecting the quality of life of T3 stage radiotherapy patients, identifying household chores ability, age standardization value, age, acceptance of post-treatment physical changes, and degree of sadness and distress as the top five core influencing factors. The clinical thresholds for each quantifiable indicator were determined through SHAP importance identification and ROC curve Youden index dual validation (household chores ability = 0.263, age standardization value/age = 0.511, acceptance of post-treatment physical changes = 1.98). At the same time, it was the first to quantitatively confirm the synergistic interaction between home chores ability and age in patients undergoing radiotherapy for abdominal malignant tumors, that is, the older the age, the more significant the positive impact of home chores ability on quality of life [39]. Subgroup analysis of tumor subtypes shows that the ranking of core influencing factors varies slightly among different abdominal malignant tumor subtypes, but their clinical thresholds are consistent with the overall sample. Moreover, the Top-5 core factors are identical to those in the overall sample, providing a scientific basis for formulating differentiated intervention strategies for different subtypes.

This study provides Level IIb evidence. The Stacking ensemble model demonstrated excellent predictive performance for T3 phase quality of life in adult patients with abdominal malignancies receiving intensity-modulated radiotherapy (IMRT) with a total dose of 45–60 Gy. The AUC of the validation set reached 0.929 (95% CI: 0.799-1.000), and within the core threshold range of 0.10-0.80, the number needed to treat (NNT) is 3, which is far below the good cost-effectiveness standard of NNT $\leqslant$ 5. Within the threshold range, it has significant clinical net benefits. The Top-5 core influencing factors and quantified clinical thresholds revealed by the SHAP analysis provide actionable core clinical targets for the precise and individualized rehabilitation interventions of T3-stage quality of life in this population. Moreover, the core features are all easily obtainable indicators in routine clinical follow-ups, without adding extra examination burdens to patients, making them suitable for rapid promotion and application in oncology radiotherapy departments at all levels of hospitals [40]. The results of this study provide direct clinical reference for the management of quality of life in patients undergoing radiotherapy for abdominal malignancies. They also lay the data foundation for subsequent large-sample, multi-center validation studies, which can offer preliminary evidence-based support for future updates to the guidelines on quality of life management for these patients.

Declarations

Abbreviations

AUC: Area Under the Curve

ANOVA: Analysis of Variance

BMI: Body Mass Index

CI: Confidence Interval

CSMS: Cancer Self-Management Scale

DCA: Decision Curve Analysis

EPV: Events Per Variable

FACT-Gv4: Functional Assessment of Cancer Therapy-General Version 4

FoP-Q-SF: Fear of Progression Questionnaire-Short Form

H-L: Hosmer-Lemeshow

ICC: Intraclass Correlation Coefficient

IMRT: Intensity-Modulated Radiation Therapy

KNN: K-Nearest Neighbor

LASSO: Least Absolute Shrinkage and Selection Operator

LOOCV: Leave-One-Out Cross Validation

LR: Logistic Regression

MAR: Missing at Random

MI: Multiple Imputation

MSE: Mean Squared Error

MSAS-Ch: Memorial Symptom Assessment Scale-Chinese Version

NNT: Number Needed to Treat

QoL: Quality of Life

RF: Random Forest

SE: Standard Error

SEL: Stacking Ensemble Learning

SHAP: SHapley Additive exPlanations

SSRS: Social Support Rating Scale

SVM: Support Vector Machine

XGBoost: Extreme Gradient Boosting

YI: Youden's Index

VIF: Variance Inflation Factor

Ethics Statement

This study was conducted in accordance with the Declaration of Helsinki (2013 revision) and approved by the Medical Ethics Committee of Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital & Institute) (Approval No.: 2026-119). The study was supported by the Scientific Research Special Fund of Beijing Wu Liande Public Welfare Foundation. Written informed consent was obtained from all patients before enrollment.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable

Data availability

Data is provided within the manuscript. More data may be provided from corresponding author on request.

Conflict of Interest

The authors declare no potential conflicts of interest. No financial or commercial relationships, or other affiliations that could be construed as influencing the research, exist between the authors and any third parties.

Funding

This work was supported by the Scientific Research Special Fund of Beijing Wu Liande Public Welfare Foundation.

Author Contributions

Conceptualization, H.L.; methodology, Y.S. and J.L.; software, Y.S.; validation, Y.S., J.L. and Y.T.; formal analysis, Y.S. and J.L.; investigation, Y.T. and W.Z.; resources, H.L.; data curation, Y.S.; writing—original draft preparation, Y.S. and J.L.; writing—review and editing, H.L.; visualization, W.Z.; supervision, H.L.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Acknowledgments

We sincerely thank all patients and their families for their participation and cooperation in this study. We are grateful to the clinicians, nurses, and medical technicians in the Department of Radiotherapy at Cancer Hospital of Dalian University of Technology (Liaoning Cancer Hospital & Institute) for their support in patient recruitment, data collection, and follow-up management. We thank the Ethics Committee for approving this study (Approval No.: 2026-119) and providing valuable ethical guidance.

We appreciate the technical support from the research team for data processing, statistical analysis, and model construction. We also thank the editors and anonymous reviewers for their constructive comments, which helped improve the quality of this manuscript. Finally, we acknowledge all researchers and institutions whose work laid the foundation for this study.

References

1. Ge, X., Liao, Z., Yuan, J., Mao, D., Li, Y., Yu, E., Wang, X., & Ding, Z. (2020). Radiotherapy-related quality of life in patients with head and neck cancers: a meta-analysis. Supportive care in cancer : official journal of the Multinational Association of Supportive Care in Cancer, 28(6), 2701–2712. https://doi.org/10.1007/s00520-019-05077-5
2. Fabian, A., Rühle, A., Domschikowski, J., Trommer, M., Wegen, S., Becker, J. N., Wurschi, G., Boeke, S., Sonnhoff, M., Fink, C. A., Käsmann, L., Schneider, M., Bockelmann, E., Treppner, M., Mehnert-Theuerkauf, A., Krug, D., Nicolay, N. H., & Young DEGRO Group (2023). Psychosocial distress in cancer patients undergoing radiotherapy: a prospective national cohort of 1042 patients in Germany. Journal of cancer research and clinical oncology, 149(11), 9017–9024. https://doi.org/10.1007/s00432-023-04837-5
3. Fabi, A., Bhargava, R., Fatigoni, S., Guglielmo, M., Horneber, M., Roila, F., Weis, J., Jordan, K., Ripamonti, C. I., & ESMO Guidelines Committee. Electronic address: clinicalguidelines@esmo.org (2020). Cancer-related fatigue: ESMO Clinical Practice Guidelines for diagnosis and treatment. Annals of oncology : official journal of the European Society for Medical Oncology, 31(6), 713–723. https://doi.org/10.1016/j.annonc.2020.02.016
4. Sharifi, F., Mohammed, E., Crump, T., & Far, B. H. (2022). Explainable Analytics to Predict the Quality of Life in Patients with Prostate Cancer from Longitudinal Data. Applied Artificial Intelligence, 36(1). https://doi.org/10.1080/08839514.2022.2055393
5. Ferrans, C. E., Zerwic, J. J., Wilbur, J. E., & Larson, J. L. (2005). Conceptual model of health-related quality of life. Journal of nursing scholarship : an official publication of Sigma Theta Tau International Honor Society of Nursing, 37(4), 336–342. https://doi.org/10.1111/j.1547-5069.2005.00058.x
6. Collins, G. S., Chester-Jones, M., Gerry, S., Ma, J., Matos, J., Sehjal, J., Tsegaye, B., & Dhiman, P. (2025). Clinical prediction models using machine learning in oncology: challenges and recommendations. BMJ oncology, 4(1), e000914. https://doi.org/10.1136/bmjonc-2025-000914
7. Mahajan, P., Uddin, S., Hajati, F., & Moni, M. A. (2023). Ensemble Learning for Disease Prediction: A Review. Healthcare (Basel, Switzerland), 11(12), 1808. https://doi.org/10.3390/healthcare11121808
8. Chae, S., Street, W. N., Ramaraju, N., & Gilbertson-White, S. (2024). Prediction of Cancer Symptom Trajectory Using Longitudinal Electronic Health Record Data and Long Sho

rt-Term Memory Neural Network. JCO clinical cancer informatics, 8, e2300039. https://doi.org/10.1200/CCI.23.00039

9. Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable Artificial Intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion. 2020;58:82-115. https://doi.org/10.1016/j.inffus.2019.12.012
10. Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in Neural Information Processing Systems 30 pp.4765−4774. Curran Associates, Inc
11. Li, M., Sun, H., Huang, Y., & Chen, H. (2024). Shapley value: From cooperative game to explainable artificial intelligence. Autonomous Intelligent Systems, 4(1), 2. https://doi.org/10.1007/s43684-023-00060-8
12. National Comprehensive Cancer Network. (2021). NCCN Radiation Therapy Compendium. Retrieved from https://www.nccn.org/compendia-templates/compendia/radiation-therapycompendium
13. Cheng, K. K., Wong, E. M., Ling, W. M., Chan, C. W., & Thompson, D. R. (2009). Measuring the symptom experience of Chinese cancer patients: a validation of the Chinese version of the memorial symptom assessment scale. Journal of pain and symptom management, 37(1), 44–57. https://doi.org/10.1016/j.jpainsymman.2007.12.019
14. Yu, C. L., Fielding, R., Chan, C. L., Tse, V. K., Choi, P. H., Lau, W. H., Choy, D. T., O, S. K., Lee, A. W., & Sham, J. S. (2000). Measuring quality of life of Chinese cancer patients: A validation of the Chinese version of the Functional Assessment of Cancer Therapy-General (FACT-G) scale. Cancer, 88(7), 1715–1727.
15. Mahendran, R., Liu, J., Kuparasundram, S., & Griva, K. (2020). Validation of the English and simplified Mandarin versions of the Fear of Progression Questionnaire - Short Form in Chinese cancer survivors. BMC psychology, 8(1), 10. https://doi.org/10.1186/s40359-020-0374-0
16. Cheng, L. L., & Sun, Y. Q. (2017). Development and validation of Cancer Patient Self-management Assessment Scale. Chinese Journal of Nursing, 52(9). https://doi.org/10.3761/j.issn.0254-1769.2017.09.014
17. Peduzzi, P., Concato, J., Kemper, E., Holford, T. R., & Feinstein, A. R. (1996). A simulation study of the number of events per variable in logistic regression analysis. Journal of c

linical epidemiology, 49(12), 1373–1379. https://doi.org/10.1016/s0895-4356(96)00236-3

18. Chen, Z., Zhao, J., Li, Y., Feng, X., Chen, Y., Li, Y., Nan, X., Liu, H., Dong, B., Shen, L., & Zhang, L. (2025). Predicting response to patients with gastric cancer via a dynamic -aware model with longitudinal liquid biopsy data. Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association, 28(5), 886–898. https://doi.org/10.1007/s10120-025-01628-4
19. Wolpert, D. H. (1992). Stacked generalization. Neural Networks, 5(2), 241-259. https://doi.org/10.1016/S0893-6080(05)80023-1
20. Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: an introduction to super learning. European journal of epidemiology, 33(5), 459–464. https://doi.org/10.1007/s10654-018-0390-z
21. Adiprakoso, D., Katsimpokis, D., Oerlemans, S., Ezendam, N. P. M., van Maaren, M. C., van Til, J. A., van der Heijden, T. G. W., Mols, F., Aben, K. K. H., Vink, G. R., Koopman, M., van de Poll-Franse, L. V., & de Rooij, B. H. (2025). Development of a prediction model for clinically-relevant fatigue: a multi-cancer approach. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation, 34(1), 231–245. https://doi.org/10.1007/s11136-024-03807-9
22. Singhal, A., Zhao, X., Wall, P., So, E., Calderini, G., Partin, A., Koussa, N., Vasanthakumari, P., Narykov, O., Zhu, Y., Jones, S. E., Abbas-Aghababazadeh, F., Nair, S. K., Bélisle-Pipon, J.-C., Jayaram, A., Parker, B. A., Yeung, K. T., Griffiths, J. I., Weil, R., ... Ideker, T. (2025). The hallmarks of predictive oncology. Cancer Discovery, 15(2), 271–285. https://doi.org/10.1158/2159-8290.CD-24-0760
23. Pergolotti, M., Deal, A. M., Williams, G. R., Bryant, A. L., Bensen, J. T., Muss, H. B., & Reeve, B. B. (2017). Activities, function, and health-related quality of life (HRQOL) of older adults with cancer. Journal of Geriatric Oncology, 8(4), 249–254. https://doi.org/10.1016/j.jgo.2017.02.009
24. Crispo, A., Pagnano, E., Bonfigli, A., Pecchia, L., Luongo, A., Porciello, G., Coluccia, S., Prete, M., Bacco, L., Vitale, S., Palumbo, E., Giaccone, P., Pica, R., Grimaldi, M., Cascella, M., Cavalcanti, E., Minopoli, A., De Laurentiis, M., Libra, M., ... Augustin, L. S. A. (2025). Adopting machine learning to predict breast cancer patients adherence with lifestyle recommendations and quality of life outcomes. Frontiers in Digital Health, 7, Article 1645233. https://doi.org/10.3389/fdgth.2025.1645233

25. Pawłowski, P., Ziętara, K. J., Milanowska, J., Aftyka, A., Sobierajski, M., Kania, Z., & S amardakiewicz, M. (2025). The Cost of Victory over Cancer: Psychosocial Dysfunction a nd Depressive Symptoms Among Polish Adolescent Cancer Survivors in the Context of Quality of Life and Psychosocial Health. Cancers, 17(24), 3916. https://doi.org/10.3390/c ancers17243916
26. Fingeret, M.C., Teo, I. and Epner, D.E. (2014), Managing body image difficulties of adul t cancer patients: Lessons from available research. Cancer, 120: 633-641. https://doi.org/ 10.1002/cncr.28469
27. Mitchell, A. J., Chan, M., Bhatti, H., Halton, M., Grassi, L., Johansen, C., & Meader, N. (2011). Prevalence of depression, anxiety, and adjustment disorder in oncological, haema tological, and palliative-care settings: A meta-analysis of 94 interview-based studies. The Lancet Oncology, 12(2), 160–174. https://doi.org/10.1016/S1470-2045(11)70002-X
28. Pintado, S., & Andrade, S. (2017). Randomized controlled trial of mindfulness program t o enhance body image in patients with breast cancer. European Journal of Integrative Me dicine, 12, 147–152. https://doi.org/10.1016/j.eujim.2017.05.009
29. Pan, X., Levin-Epstein, R., Huang, J., Ruan, D., King, C. R., Kishan, A. U., Steinberg, M. L., & Qi, X. S. (2020). Dosimetric predictors of patient-reported toxicity after prostate st ereotactic body radiotherapy: Analysis of full range of the dose-volume histogram using ensemble machine learning. Radiotherapy and oncology : journal of the European Societ y for Therapeutic Radiology and Oncology, 148, 181–188. https://doi.org/10.1016/j.rado nc.2020.04.013
30. Grassi, L., Caruso, R., Riba, M. B., Lloyd-Williams, M., Kissane, D., Rodin, G., McFarla nd, D., Campos-Ródenas, R., Zachariae, R., Santini, D., Ripamonti, C. I., & ESMO Guid elines Committee. Electronic address: clinicalguidelines@esmo.org (2023). Anxiety and depression in adult cancer patients: ESMO Clinical Practice Guideline. ESMO open, 8(2), 101155. https://doi.org/10.1016/j.esmoop.2023.101155
31. Vickers, A. J., van Calster, B., & Steyerberg, E. W. (2019). A simple, step-by-step guide to interpreting decision curve analysis. Diagnostic and prognostic research, 3, 18. https:// doi.org/10.1186/s41512-019-0064-7
32. Kang, D., Kim, H., Cho, J., Kim, Z., Chung, M., Lee, J. E., Nam, S. J., Kim, S. W., Yu, J., Chae, B. J., Ryu, J. M., & Lee, S. K. (2023). Prediction Model for Postoperative Quali ty of Life Among Breast Cancer Survivors Along the Survivorship Trajectory From Pretr eatment to 5 Years: Machine Learning-Based Analysis. JMIR public health and surveilla nce, 9, e45212. https://doi.org/10.2196/45212

33. Dhiman, P., Ma, J., Navarro, C. A., Speich, B., Bullock, G., Damen, J. A., Kirtley, S., Hooft, L., Riley, R. D., Van Calster, B., Moons, K. G. M., & Collins, G. S. (2021). Reporting of prognostic clinical prediction models based on machine learning methods in oncology needs to be improved. Journal of clinical epidemiology, 138, 60–72. https://doi.org/10.1016/j.jclinepi.2021.06.024
34. Van der Heijden, T. G. W., de Ligt, K. M., Hubel, N. J., van der Mierden, S., Holzner, B., van de Poll-Franse, L. V., de Rooij, B. H., & EORTC Quality of Life Group (2025). Exploring the role of health-related quality of life measures in predictive modelling for oncology: a systematic review. Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation, 34(2), 305–323. https://doi.org/10.1007/s11136-024-03820-y
35. Von Elm, E., Altman, D. G., Egger, M., Pocock, S. J., Gøtzsche, P. C., Vandenbroucke, J. P., & STROBE Initiative (2008). The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Journal of clinical epidemiology, 61(4), 344–349. https://doi.org/10.1016/j.jclinepi.2007.11.008
36. Riley, R. D., Ensor, J., Snell, K. I. E., Debray, T. P. A., Altman, D. G., Moons, K. G. M., & Collins, G. S. (2016). External validation of clinical prediction models using big datasets from e-health records or IPD meta-analysis: Opportunities and challenges. BMJ, 353, i3140. https://doi.org/10.1136/bmj.i3140
37. OCEBM Levels of Evidence Working Group. (2011). The Oxford 2011 Levels of Evidence. Oxford Centre for Evidence-Based Medicine. https://www.cebm.ox.ac.uk/resources/levels-of-evidence/ocebm-levels-of-evidence
38. Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: a novel method for evaluating prediction models. Medical decision making : an international journal of the Society for Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361
39. Molnar, C. (2022). Interpretable machine learning: A guide for making black box models explainable (2nd ed.). Christoph Molnar.
40. Beam, A. L., & Kohane, I. S. (2018). Big Data and Machine Learning in Health Care. JAMA, 319(13), 1317–1318. https://doi.org/10.1001/jama.2017.18391

Publisher's Note

The views expressed in this article are solely those of the authors and do not necessarily reflect the official policies or positions of their affiliated institutions, the funding agency, the publisher, editors, or reviewers. The publisher assumes no responsibility for any errors or omissions in the content, nor does it guarantee or endorse any products or claims mentioned herein.