

Establishment of a Multimodal Prognostic Prediction Model for Multiple Myeloma Patients Based on Radiomics and Clinical Features: A Retrospective Cohort Study

Radyomik ve Klinik Özelliklere Dayalı Multimodal Prognostik Tahmin Modelinin Multipl Miyelom Hastalarında Geliştirilmesi: Retrospektif Bir Kohort Çalışması

Shiqi Sun^{1†}, Zejing Huang^{1†}, Yunlong Tang¹, Weiying Gu², Leilei Wu¹, Fujun Shen³

¹Affiliated Hospital Six of Nantong University, Yancheng Third People's Hospital, Department of Hematology, Yancheng, P.R. China

²The Third Affiliated Hospital of Soochow University, Department of Hematology, Suzhou, P.R. China

³Affiliated Hospital Six of Nantong University, Yancheng Third People's Hospital, Department of Oncology, Yancheng, P.R. China

[†]These authors contributed equally to this study.

Abstract

Objective: We aimed to develop a survival prediction system integrating radiomics and clinical features for newly diagnosed multiple myeloma (MM) and to compare the performance of different feature sets and algorithms in the early prediction of progression-free survival (PFS).

Materials and Methods: This study retrospectively included 300 MM patients between June 2022 and June 2024, with their baseline positron emission tomography, computed tomography, and magnetic resonance imaging radiomics features and clinical variables collected. Following construction of a radiomics-based risk score (Rad-score), seven machine learning-based survival models were established using the clinical feature set, the image feature set, and the integrated feature set.

Results: The fusion feature set-based gradient boosting model (GBM) showed numerically favorable overall performance in predicting 12-month PFS, suggesting that the integration of radiomics and clinical variables may provide complementary predictive information for early risk stratification. The model effectively distinguished high-, intermediate-, and low-risk patients (log-rank $p < 0.001$), and calibration curve analysis and decision curve analysis revealed favorable calibration and high clinical net benefits. Shapley additive explanations analysis showed that the Rad-score was one of the most important features in the model, indicating that radiomics information may contribute prognostic value within the integrated feature space. β 2-microglobulin, age, lactate dehydrogenase, blood calcium, platelet count, and hemoglobin were also identified as key contributing features.

Öz

Amaç: Bu çalışmada, yeni tanı almış multipl miyelom (MM) hastalarında radyomik ve klinik özellikleri bütünleştiren bir sağkalım tahmin sistemi geliştirilmesi ve progresyonsuz sağkalımın (PFS) erken tahmininde farklı özellik setleri ile algoritmaların performanslarının karşılaştırılması amaçlandı.

Gereç ve Yöntemler: Bu retrospektif çalışmaya Haziran 2022 ile Haziran 2024 tarihleri arasında 300 MM hastası dahil edildi. Hastaların başlangıç pozitron emisyon tomografisi, bilgisayarlı tomografi ve manyetik rezonans görüntüleme radyomik özellikleri ile klinik değişkenleri toplandı. Radyomik tabanlı bir risk skoru (Rad-skoru) oluşturulduktan sonra klinik özellik seti, görüntüleme özellik seti ve bütünleşik özellik seti kullanılarak makine öğrenmesine dayalı yedi sağkalım modeli geliştirildi.

Bulgular: Birleşik özellik setine dayalı gradyan artırma modeli (GBM), 12 aylık PFS'nin tahmininde sayısal olarak daha iyi bir genel performans gösterdi. Bu bulgu, radyomik ve klinik değişkenlerin bütünleştirilmesinin erken risk sınıflandırması açısından tamamlayıcı öngörü bilgileri sağlayabileceğini düşündürdü. Model; yüksek, orta ve düşük riskli hastaları etkili biçimde ayırt etti (log-rank $p < 0.001$). Kalibrasyon eğrisi ve karar eğrisi analizleri, modelin uygun kalibrasyona ve yüksek klinik net faydaya sahip olduğunu gösterdi. Shapley katkı açıklamaları analizi, Rad-skorunun modeldeki en önemli özelliklerden biri olduğunu ortaya koyarak radyomik bilgilerin bütünleşik özellik alanında prognostik değer sağlayabileceğini gösterdi. β 2-mikroglobulin, yaş, laktat dehidrogenaz, kan kalsiyum düzeyi, trombosit sayısı ve hemoglobin de modele önemli katkı sağlayan özellikler olarak belirlendi.



Address for Correspondence/Yazışma Adresi: Yunlong Tang, M.D., Affiliated Hospital Six of Nantong University, Yancheng Third People's Hospital, Department of Hematology, Yancheng, P.R. China
E-mail: tangylytph@univ-edu.org ORCID: orcid.org/0009-0007-6937-2602

Address for Correspondence/Yazışma Adresi: Fujun Shen, M.D., Affiliated Hospital Six of Nantong University, Yancheng Third People's Hospital, Department of Oncology, Yancheng, P.R. China
E-mail: shenfjytp@sdsc.cn ORCID: orcid.org/0009-0006-9956-4739

Received/Geliş tarihi: April 20, 2026

Accepted/Kabul tarihi: July 13, 2026

Epub: July 16, 2026



©Copyright 2026 by Turkish Society of Hematology Turkish Journal of Hematology, Published by Galenos Publishing House.
Licensed under a Creative Commons Attribution-NonCommercial (CC BY-NC-ND) 4.0 International License.

Abstract

Conclusion: The integration of radiomics and clinical variables may improve early PFS prediction in MM patients. The GBM using fused features showed relatively good discriminative ability, potential clinical applicability, and favorable interpretability.

Keywords: Machine learning, Multiple myeloma, Prediction, Progression-free survival, Radiomics

Öz

Sonuç: Radyomik ve klinik değişkenlerin bütünleştirilmesi, MM hastalarında erken PFS tahminini iyileştirebilir. Birleşik özellikleri kullanan GBM, görece iyi ayırt edici performans, potansiyel klinik uygulanabilirlik ve uygun yorumlanabilirlik gösterdi.

Anahtar Sözcükler: Makine öğrenmesi, Multipl miyelom, Tahmin, Progresyonsuz sağkalım, Radyomik

Introduction

Multiple myeloma (MM) is a hematological malignancy characterized by clonal plasma cell proliferation [1]. Although novel agents and hematopoietic stem cell transplantation have improved survival, prognosis remains highly heterogeneous, ranging from durable remission to early progression and ultra-high-risk disease [2]. Current risk stratification systems, such as the revised International Staging System (R-ISS), incorporate cytogenetics and lactate dehydrogenase (LDH), but still cannot fully explain prognostic differences in real-world patients, especially those at intermediate risk [3]. Therefore, more comprehensive prediction strategies are needed.

Imaging is essential for the diagnosis, staging, and response assessment of MM. Whole-body low-dose computed tomography (CT), magnetic resonance imaging (MRI), and ¹⁸F-fluorodeoxyglucose positron emission tomography (FDG PET)/CT provide complementary information on osteolytic lesions, bone marrow infiltration, extramedullary disease, and residual lesions [4,5]. Previous studies have also demonstrated the complementary value of MRI and PET/CT in evaluating disease burden and treatment response [6]. However, conventional imaging assessment mainly relies on visual interpretation and limited qualitative or semi-quantitative parameters, which may not fully capture the prognostic information contained in medical images.

Radiomics and artificial intelligence provide new approaches for extracting prognosis-related imaging phenotypes from high-dimensional medical images. Recent studies have developed MRI- or PET/CT-based radiomics models to predict cytogenetic abnormalities, minimal residual disease, and survival outcomes in MM, suggesting that radiomics may improve traditional prognostic stratification [7]. However, most studies have focused on a single imaging modality, used relatively small single-center cohorts, or lacked systematic comparison across different survival models [8,9]. Moreover, clinical machine learning models for overall survival or progression-free survival (PFS) prediction have often relied mainly on clinical and laboratory variables, with limited integration of radiomics, psychological status, and quality-of-life indicators [10,11].

To address these limitations, this study enrolled 300 newly diagnosed MM patients and integrated PET, CT, and MRI radiomics features to construct a radiomics-based risk score (Rad-score). The Rad-score was combined with clinical variables to develop fusion feature sets, and seven survival models were compared for predicting 12-month PFS. Model performance was evaluated using time-dependent receiver operating characteristic (ROC) curve analysis, Kaplan-Meier analysis, calibration curves, decision curve analysis (DCA), and Shapley additive explanations (SHAP) analysis. Thus, this study aimed to develop a multimodal, interpretable, and clinically applicable survival prediction strategy for MM.

Materials and Methods

Study Design and Subjects

This retrospective cohort study included 300 patients with MM diagnosed and treated at our hospital from June 2022 to June 2024. All patients underwent pretreatment PET/CT and MRI examinations and had the necessary available baseline clinical, laboratory, imaging, and follow-up data for model development and validation. The study was approved by the Ethics Committee of the Affiliated Hospital Six of Nantong University, Yancheng Third People's Hospital (no: YTP2022-026, approved on May 11, 2022), and written informed consent was obtained from all patients.

Inclusion and Exclusion Criteria

Patients were included if they were aged ≥ 18 years, were clinically diagnosed with MM according to International Myeloma Working Group criteria, had undergone pretreatment PET/CT and MRI examinations, and had the essential clinical, imaging, and follow-up data required for model development and validation. Patients were excluded if they had received prior chemotherapy or radiotherapy, had other malignancies, were pregnant or lactating, had poor-quality images, or were participating in any clinical studies.

Treatment Methods

All patients received guideline-based anti-myeloma treatment according to current diagnostic and treatment guidelines. Treatment regimens were selected based on transplant eligibility,

disease status, baseline risk stratification, organ function, and physician judgment. To account for potential treatment-related heterogeneity, detailed treatment information was collected and incorporated as candidate clinical variables. These variables included the type of induction regimen, such as bortezomib and dexamethasone (Vd), lenalidomide and dexamethasone (Rd), or bortezomib, lenalidomide, and dexamethasone (RVd); the number of treatment cycles; use of consolidation therapy; receipt of autologous hematopoietic stem cell transplantation; and use of novel agents. Drug types and doses were also reviewed from medical records when available. These treatment-related variables were included in model development to reduce the potential confounding effect of treatment differences on PFS.

Outcome Measures and Follow-Up

PFS was the primary outcome, defined as the time from diagnosis to the first occurrence of disease progression, clinical relapse, or death from any cause. Patients without these events by the end of the follow-up period were censored at the date of last follow-up. In this study, 12-month PFS was selected as the predefined time point for model evaluation because progression within the first year reflects early disease progression, which may indicate aggressive disease biology, suboptimal treatment response, and poorer prognosis in MM. In addition, all enrolled patients had a potential follow-up window of at least 12 months, supporting stable assessment at this time point.

Disease progression and clinical relapse after complete remission were counted as PFS events according to established criteria [12]. Disease progression was defined as any of the following changes from the lowest recorded value: an increase of $\geq 25\%$ in serum M-protein, urine M-protein, the difference between involved and uninvolved serum free light chain, or bone marrow plasma cells, with the required absolute increases; the appearance of new soft tissue plasmacytoma lesions or $\geq 50\%$ enlargement of existing measurable lesions; or a $\geq 50\%$ increase in circulating plasma cells when they were the only measurable lesions. Clinical relapse was defined as the appearance of new bone lesions or soft tissue plasmacytomas, clear progression of existing lesions, hypercalcemia, a decrease in hemoglobin of ≥ 20 g/L unrelated to treatment or non-MM factors, an increase in serum creatinine of ≥ 176.8 $\mu\text{mol/L}$ after treatment initiation, or hyperviscosity related to serum M-protein.

All progression and relapse events were independently assessed by two experienced hematologists and disagreements were resolved by discussion. Patients were followed through outpatient visits, inpatient re-examinations, and telephone interviews, with regular evaluation of laboratory indicators, imaging findings, and disease status. The follow-up cutoff date was June 30, 2025, ensuring a potential follow-up window of at least 12 months for all enrolled patients. During follow-up, PFS events occurred for 123 patients in the overall cohort, including 86 patients in the training cohort and 37 patients in the test cohort.

Dataset Division and Prevention of Information Leakage

The whole cohort was randomly divided into a training cohort and an independent test cohort at a ratio of 7:3 before any data-driven analysis. The training cohort included 210 patients, of whom 86 had experienced PFS events, whereas the independent test cohort included 90 patients, of whom 37 had experienced PFS events. The event proportion was comparable between the two cohorts, supporting the representativeness of the split. All preprocessing and model-development procedures were performed in a strictly training-set-only manner. Specifically, missing-value imputation models, normalization parameters, variance filtering, correlation filtering, radiomics feature selection, Rad-score construction, and hyperparameter tuning were fitted exclusively in the training cohort. The derived parameters, retained feature set, and final model coefficients or model structures were then directly applied to the independent test cohort. No test-set information was used for feature selection, parameter optimization, model fitting, or threshold determination.

To reduce overfitting, several safeguards were used. First, dimensionality reduction was performed before model training; second, radiomics features were screened by reproducibility, variance, correlation, and least absolute shrinkage and selection operator (LASSO)-Cox selection; third, hyperparameters were optimized by cross-validation within the training cohort only; and fourth, the final model was evaluated once in an independent test cohort.

Overall Study Workflow

The overall workflow of patient enrollment, image acquisition, radiomics feature extraction, feature selection, model construction, validation, and interpretation is summarized in Figure 1.

Collection and Preprocessing of Clinical Variables

Baseline clinical variables included demographic data, Eastern Cooperative Oncology Group (ECOG) score, psychological and quality-of-life indicators, laboratory results, high-risk cytogenetic abnormalities, and treatment-related factors. Psychological indicators included anxiety and depression scores. Anxiety and depression were assessed using the Hospital Anxiety and Depression Scale-Anxiety subscale (HADS-A) and Hospital Anxiety and Depression Scale-Depression subscale (HADS-D), respectively. Quality of life was evaluated using the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30 (EORTC QLQ-C30). All psychological and quality-of-life assessments were completed at baseline before treatment, and scores were calculated according to the standard scoring instructions for each questionnaire. Candidate variables were assessed for completeness and multicollinearity. Variables with $>20\%$ missing data were excluded, while variables

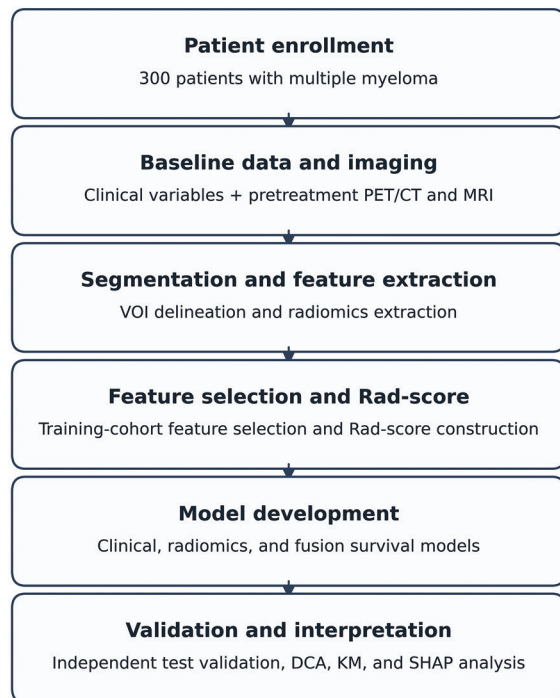


Figure 1. Overall study workflow.

PET/CT: Positron emission tomography/computed tomography; MRI: magnetic resonance imaging; VOI: volume of interest; Rad-score: radiomics-based risk score; DCA: decision curve analysis; KM: Kaplan-Meier; SHAP: Shapley additive explanations.

with $\leq 20\%$ missing data were imputed using multiple chained equations based on the training set. The imputation model derived from the training cohort was subsequently applied to the test cohort, and no information from the test cohort was used during imputation model fitting. A total of 23 clinically relevant and stable variables were retained for modeling.

Multimodal Image Data Acquisition and Lesion Delineation

All patients underwent pretreatment ^{18}F -FDG PET/CT and MRI examinations. PET/CT images were obtained using a Biograph mCT scanner (Siemens, Munich, Germany). According to the routine institutional protocol, PET imaging was performed approximately 60 min after intravenous injection of ^{18}F -FDG. MRI was performed using a 3.0 T Magnetom Skyra scanner (Siemens). The MRI protocol entailed conventional anatomical and functional sequences, including T1-weighted imaging, T2-weighted imaging, short tau inversion recovery, and diffusion-weighted imaging. Digital Imaging and Communications in Medicine (DICOM) images were imported to ITK-SNAP software for manual delineation of volumes of interest. For patients with multiple lesions, target lesions were selected according to lesion visibility, measurable size, image quality, and consistency across PET/CT and MRI. Lesions with unclear boundaries, severe artifacts, or insufficient size for reliable segmentation were excluded. When multiple eligible lesions were present,

the representative lesion with the largest measurable size or highest metabolic activity was preferentially selected. Lesions were independently segmented by two experienced radiologists blinded to clinical data, and discrepancies were reviewed by a senior radiologist. Segmentation was performed with reference to both anatomical abnormalities and PET metabolic uptake to ensure consistency of lesion localization across modalities.

Radiomics Feature Extraction and Stability Screening

Before feature extraction, PET, CT, and MRI images were preprocessed using a standardized workflow to reduce inter-modality variability. All images were resampled to a uniform voxel spacing before feature extraction. PET images were converted to standardized uptake values, CT images were analyzed in Hounsfield units, and MRI signal intensities were normalized to reduce scanner- and sequence-related variability. Gray-level discretization was performed using consistent settings across all patients. After image preprocessing, radiomics features were extracted from PET, CT, and MRI images using the open-source Pyradiomics package. Extracted features included shape, histogram, and texture features. Feature stability was assessed using the intraclass correlation coefficient (ICC), and features with ICC of >0.75 were retained. To avoid data leakage, subsequent low-variance filtering and correlation-based feature reduction were performed only in the training cohort. The same retained feature set was then applied to the test cohort. After this process, 146 stable radiomics features were retained for subsequent analysis.

The radiomics feature-selection strategy followed a predefined sequential procedure. First, features with insufficient segmentation reproducibility were removed according to the ICC threshold. Second, low-variance features were excluded in the training cohort. Third, highly correlated features were reduced using correlation analysis to decrease redundancy. Fourth, the remaining radiomics features were entered into LASSO-Cox regression for supervised dimensionality reduction. Importantly, all data-driven feature filtering steps after the initial reproducibility assessment were fitted only in the training cohort, and the identical retained feature set was subsequently applied to the test cohort.

Establishment of the Rad-Score

The 146 retained features were subjected to LASSO-Cox regression for further dimensionality reduction. LASSO-Cox feature selection was performed exclusively in the training cohort. The optimal penalty parameter $\lambda_{\min}$ was selected through 10-fold cross-validation within the training cohort, and features with non-zero coefficients were used to construct the Rad-score as follows:

$$\text{Rad-score} = \sum_{i=1}^k (\beta_i \times \text{Feature}_i)$$

Here, β_i represents the non-zero regression coefficient in the LASSO-Cox model, Feature_i is the corresponding radiomics feature, and k represents the number of features finally retained. The Rad-score formula established in the training cohort was then directly applied to the test cohort for model evaluation. This procedure ensured that radiomics feature selection and Rad-score construction were fully nested within the training cohort, thereby reducing the risk of information leakage. As a unified multimodal imaging phenotype, the Rad-score comprehensively reflected the morphological, metabolic, and textural information in PET, CT and MRI images.

Construction of Fusion Feature Set

To reduce dimensionality and improve model stability, the Rad-score was combined with the 23 clinical variables to form a fusion feature set containing 24 features. All variables were standardized using Z-score normalization before model training. The mean and standard deviation used for Z-score normalization were calculated from the training cohort and then applied unchanged to the test cohort.

Establishment and Assessment of Prognostic Models

Prognostic models were developed using three feature sets: clinical features, radiomics features, and fusion features. Seven survival models were constructed for each feature set, including the Cox, LASSO-Cox, gradient boosting (GB)-Cox, CoxBoost, gradient boosting model (GBM), random survival forest, and support vector censored regression (SVCR) models. Hyperparameters were optimized using fivefold cross-validation in the training set, and model performance was evaluated in the independent test set.

For machine-learning survival models, hyperparameter optimization was performed using a grid search combined with fivefold cross-validation within the training cohort. The candidate hyperparameter space was prespecified before model evaluation. For tree-based models, the tuned parameters included the number of trees, maximum tree depth, learning rate, minimum samples per terminal node, and/or subsampling parameters, as appropriate for each algorithm. For the CoxBoost model, the number of boosting steps and penalty-related parameters were tuned. For SVCR, the kernel-related and regularization parameters were optimized. The hyperparameter combination with the best cross-validated performance in the training cohort was selected, after which the final model was refitted using the full training cohort and evaluated in the independent test cohort only.

Model Performance Assessment and Interpretability Analysis

Model performance was assessed using time-dependent ROC curves and 12-month area under the curve (AUC), sensitivity, specificity, accuracy, precision, and F1 score values. To provide a more comprehensive evaluation of model performance, the concordance index (C-index) and Brier score were also calculated. The 95% confidence intervals (CIs) for performance metrics were estimated using bootstrap resampling. The model with the best descriptive performance was further evaluated by Kaplan-Meier survival analysis, calibration curve, and DCA. SHAP analysis was used to assess feature importance and improve model interpretability.

Statistical Analysis

Statistical analyses were performed using IBM SPSS Statistics 26.0 (IBM Corp., Armonk, NY, USA) and Python 3.9 (Python Software Foundation, Wilmington, DE, USA). Continuous variables were expressed as mean $\pm$ standard deviation or median with interquartile range (IQR), depending on the distribution, and compared using t-tests or Mann-Whitney U tests. Categorical variables were expressed as frequencies and percentages and compared using chi-square or Fisher exact tests. A two-sided value of $p < 0.05$ was considered statistically significant.

Results

Baseline Characteristics of Patients

A total of 300 patients with MM were included. The mean age was 64.9 ± 8.1 years, 55.7% were male, and the mean body mass index value was 23.5 ± 3.5 kg/m². The median ECOG score was 1 (IQR: 0-2). High-risk cytogenetic abnormalities were present in 26.3% of patients. Key disease-related indicators, including ISS stage, R-ISS stage, β_2 -microglobulin, LDH, hemoglobin, blood calcium, and creatinine, were also recorded and are summarized in Table 1. Vd, Rd, and RVd regimens accounted for 47.0%, 28.0%, and 25.0% of treatments, respectively. Autologous hematopoietic stem cell transplantation was performed for 64.7% of patients and 47.7% received novel agents. No significant differences were observed between the training and test cohorts in baseline clinical characteristics, disease-related indicators, or treatment-related variables, indicating that the two cohorts were generally comparable. Overall, the cohort showed heterogeneity in clinical and treatment characteristics, supporting subsequent prognostic model development.

Time-Dependent AUC Analysis for Different Feature Sets

Seven survival models were developed for each of the clinical, radiomics, and fusion feature sets. Time-dependent ROC analysis showed that the clinical and fusion feature sets had similar predictive performance, whereas the radiomics feature set showed lower AUC values across most time points (Figures

Table 1. Baseline clinical and treatment-related characteristics of the training and test cohorts.

Variables	Overall cohort (n=300)	Training cohort (n=210)	Test cohort (n=90)	p
Age, years	64.9±8.1	64.8±8.2	65.1±7.9	0.769
Sex, male, n (%)	167 (55.7)	117 (55.7)	50 (55.6)	0.919
BMI, kg/m ²	23.5±3.5	23.4±3.5	23.7±3.4	0.493
ECOG score	1 (0-2)	1 (0-2)	1 (0-2)	0.866
ISS stage, I/II/III, n	71/132/97	49/94/67	22/38/30	0.891
R-ISS stage, I/II/III, n	59/162/79	41/114/55	18/48/24	0.969
β2-microglobulin, mg/L	5.1 (3.4-7.9)	5.2 (3.4-8.0)	5.0 (3.3-7.8)	0.713
Lactate dehydrogenase, U/L	225 (184-292)	226 (185-294)	222 (181-289)	0.671
Hemoglobin, g/L	96.8±18.6	96.5±18.9	97.6±17.9	0.639
Blood calcium, mmol/L	2.36±0.28	2.36±0.29	2.35±0.27	0.780
Creatinine, μmol/L	85 (65-122)	86 (66-124)	84 (63-119)	0.604
High-risk cytogenetic abnormalities, n (%)	79 (26.3)	55 (26.2)	24 (26.7)	0.954
HADS-A score	7.8±3.2	7.9±3.3	7.6±3.1	0.463
HADS-D score	7.5±3.1	7.6±3.2	7.3±3.0	0.449
EORTC QLQ-C30 global health status score	58.4±15.7	58.0±15.9	59.2±15.4	0.546
Induction regimen, n (%)				0.990
Vd	141 (47.0)	99 (47.1)	42 (46.7)	
Rd	84 (28.0)	59 (28.1)	25 (27.8)	
RVd	75 (25.0)	52 (24.8)	23 (25.6)	
Number of treatment cycles	6 (4-8)	6 (4-8)	6 (4-8)	0.881
Consolidation therapy, n (%)	143 (47.7)	99 (47.1)	44 (48.9)	0.880
Autologous hematopoietic stem cell transplantation, n (%)	194 (64.7)	136 (64.8)	58 (64.4)	0.937
Use of novel agents, n (%)	143 (47.7)	100 (47.6)	43 (47.8)	0.946
PFS events, n (%)	123 (41.0)	86 (41.0)	37 (41.1)	0.918

BMI: Body mass index; ECOG: Eastern Cooperative Oncology Group score; ISS: International Staging System; R-ISS: revised International Staging System; HADS-A: Hospital Anxiety and Depression Scale-Anxiety subscale; HADS-D: Hospital Anxiety and Depression Scale-Depression subscale; EORTC QLQ-C30: European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire-Core 30; Vd: bortezomib and dexamethasone; Rd: lenalidomide and dexamethasone; RVd: bortezomib, lenalidomide, and dexamethasone; PFS: progression-free survival.

2A-2C). Although the fusion feature set showed slightly higher AUC values at some time points, the differences were limited. Therefore, further evaluation at a fixed clinically relevant time point was performed.

Comparison of Model Performance at 12 Months

At the 12-month time point, a clinically meaningful time point corresponding to the highest relapse density [13], 21 models based on clinical, radiomics, and fusion feature sets were compared in the independent test cohort. The test cohort included 90 patients, among whom 37 experienced PFS events within 12 months. Model performance was evaluated using AUC, sensitivity, specificity, accuracy, precision, F1 score, the Harrell C-index, and the 12-month Brier score.

Among all candidate models, the fusion feature set-based GBM showed the best overall descriptive performance, with

the highest F1 score of 0.825 and a high AUC of 0.892. In the independent test cohort, this model achieved sensitivity of 0.892, specificity of 0.811, accuracy of 0.844, precision of 0.767, and F1 score of 0.825. The 12-month AUC of the fusion GBM was 0.892 (95% CI: 0.818-0.966) (Table 2). Survival-specific evaluation further showed a Harrell C-index value of 0.804 (95% CI: 0.745-0.860) and 12-month Brier score of 0.143 (95% CI: 0.107-0.180), indicating good discriminative ability and acceptable prediction error.

Although the clinical Cox model showed a slightly higher AUC, the fusion GBM had the highest F1 score and a favorable balance between sensitivity and specificity. Therefore, the fusion feature set-based GBM was selected as the candidate model with the best overall descriptive performance for predicting 12-month progression and was used for subsequent risk stratification, calibration, DCA, and SHAP interpretation.

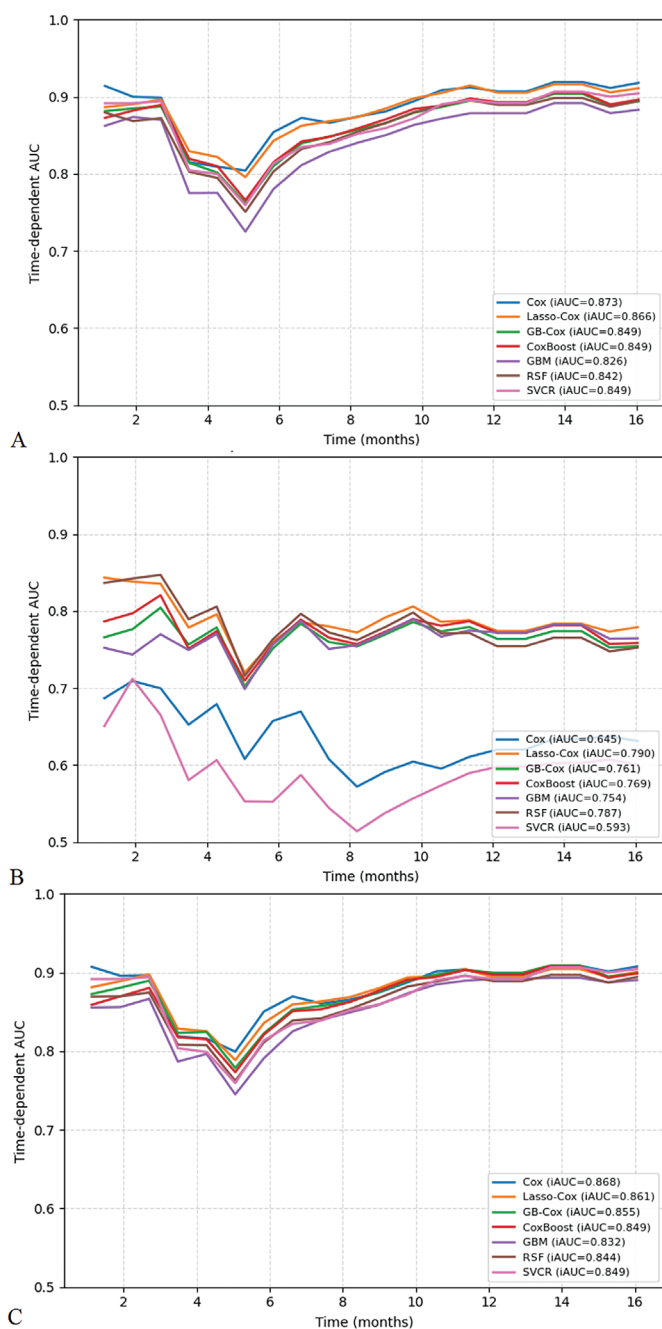


Figure 2. Time-dependent receiver operating characteristic curve performance of seven machine learning-based survival models for three feature sets: A) clinical feature set, B) radiomics feature set, C) fusion feature set.

AUC: Area under the curve; Lasso: least absolute shrinkage and selection operator; GB: gradient boosting; GBM: gradient boosting model; RSF: random survival forest; SVCR: support vector censored regression.

Risk Stratification Ability of the Fusion Feature Set-Based GBM

Patients in the independent test set were divided into low-, intermediate-, and high-risk groups according to tertiles of the risk scores generated by the fusion feature set-based GBM.

Kaplan-Meier analysis showed clear separation of PFS curves among the three groups (Figure 3). The high-risk group had the poorest PFS, whereas the low-risk group had the most favorable outcome. The difference was statistically significant (log-rank $p < 0.001$), indicating good risk stratification ability.

Assessment of Calibration Ability of the Fusion Feature Set-Based GBM

The 12-month calibration curve showed good agreement between predicted and observed PFS probabilities in the test set (Figure 4). The curve was close to the ideal reference line across most risk intervals, suggesting satisfactory calibration and reliable probability estimation.

Clinical Net Benefit of the Fusion Feature Set-Based GBM Evaluated by DCA

DCA showed that the fusion feature set-based GBM provided greater net benefit than the “treat-all” and “treat-none” strategies across a threshold probability range of approximately 0.10-0.70 (Figure 5). It also showed higher descriptive net benefit than models based on clinical or radiomics features alone across most threshold ranges, suggesting its potential value for individualized risk assessment.

Feature Importance Analysis of the Fusion Feature Set-Based GBM

SHAP analysis showed that the Rad-score was among the most important predictors, indicating that radiomics provided additional prognostic information (Figures 6A and 6B). Important clinical predictors included $\beta 2$ -microglobulin, age, LDH, blood calcium, platelet count, and hemoglobin, which are recognized prognostic factors in MM [14,15]. High-risk cytogenetic abnormalities, renal function, performance status, and emotional and quality-of-life indicators also contributed to model prediction. These findings suggest that the fusion model integrated imaging and clinical information for multidimensional prognostic assessment.

Discussion

The present study explored the integration of PET, CT, and MRI radiomics features with 23 clinical variables and compared seven machine learning-based survival models. The fusion feature set-based GBM achieved the highest F1 score and showed balanced overall performance among the evaluated models, although some clinical models had comparable or slightly higher AUC values. SHAP analysis further showed that the Rad-score ranked among the important predictors, suggesting that multimodal radiomics may provide complementary prognostic information beyond conventional clinical factors.

Previous studies have reported the potential value of radiomics in MM prognosis. Guinea-Pérez et al. [16] constructed a PET/CT-based

Table 2. Comparison of model performance at 12 months, ranked by F1 values.

Ranking	Feature set	Model	AUC	Sensitivity	Specificity	Accuracy	Precision	F1
1	Fusion	GBM	0.892 (0.818-0.966)	0.892	0.811	0.844	0.767	0.825
2	Clinical	RSF	0.889 (0.814-0.964)	0.973	0.717	0.822	0.706	0.818
3	Clinical	Cox	0.907 (0.838-0.976)	0.838	0.849	0.844	0.795	0.816
4	Clinical	GB-Cox	0.892 (0.818-0.966)	0.865	0.811	0.833	0.762	0.810
5	Fusion	RSF	0.889 (0.814-0.964)	0.865	0.811	0.833	0.762	0.810
6	Clinical	GBM	0.879 (0.801-0.957)	0.919	0.755	0.822	0.723	0.810
7	Fusion	GB-Cox	0.900 (0.829-0.971)	0.838	0.830	0.833	0.775	0.805
8	Clinical	LASSO-Cox	0.905 (0.835-0.975)	0.946	0.717	0.811	0.700	0.805
9	Fusion	CoxBoost	0.898 (0.826-0.970)	0.865	0.793	0.822	0.744	0.800
10	Fusion	SVCR	0.892 (0.818-0.966)	0.757	0.906	0.844	0.849	0.800
11	Clinical	SVCR	0.892 (0.818-0.966)	0.757	0.906	0.844	0.849	0.800
12	Clinical	CoxBoost	0.893 (0.819-0.967)	0.865	0.793	0.822	0.744	0.800
13	Fusion	Cox	0.897 (0.825-0.969)	0.757	0.887	0.833	0.824	0.789
14	Fusion	LASSO-Cox	0.894 (0.821-0.967)	0.892	0.717	0.789	0.688	0.777
15	Radiomics	LASSO-Cox	0.774 (0.672-0.876)	0.784	0.755	0.767	0.691	0.734
16	Radiomics	GBM	0.772 (0.670-0.874)	0.703	0.811	0.767	0.722	0.712
17	Radiomics	GB-Cox	0.764 (0.660-0.868)	0.676	0.830	0.767	0.735	0.704
18	Radiomics	CoxBoost	0.772 (0.670-0.874)	0.649	0.849	0.767	0.750	0.696
19	Radiomics	RSF	0.754 (0.649-0.859)	0.622	0.830	0.744	0.719	0.667
20	Radiomics	Cox	0.620 (0.501-0.739)	0.703	0.547	0.611	0.520	0.598
21	Radiomics	SVCR	0.598 (0.478-0.718)	0.351	0.868	0.656	0.650	0.456

AUC: Area under the curve; GBM: gradient boosting model; RSF: random survival forest; GB: gradient boosting; LASSO: least absolute shrinkage and selection operator; SVCR: support vector censored regression.

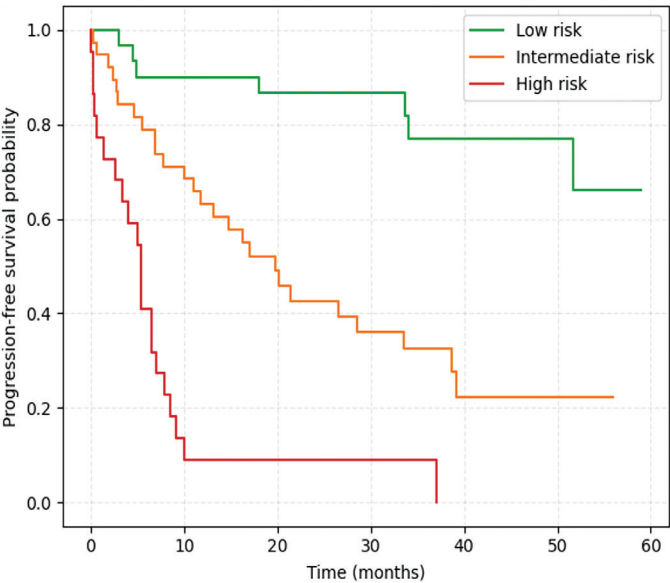


Figure 3. Risk stratification performance of the fusion feature set-based gradient boosting model (GBM) in the test set. Red, orange, and green lines indicate the high-, intermediate-, and low-risk groups, respectively.

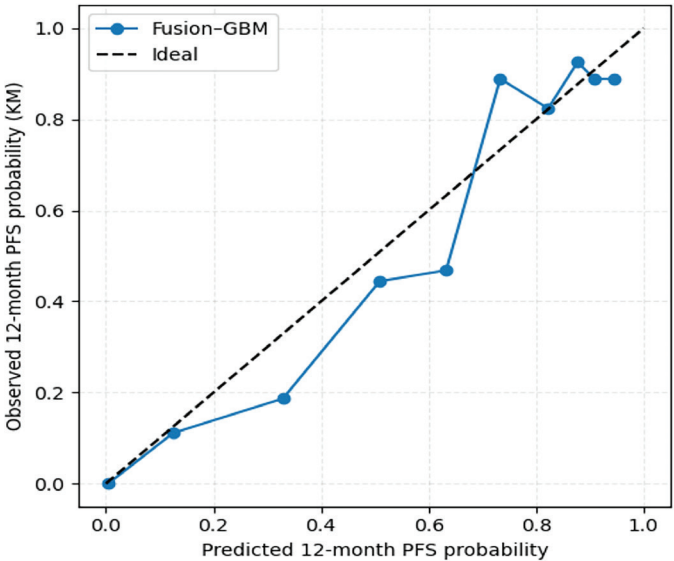


Figure 4. Twelve-month calibration curve of the fusion feature set-based GBM on the test set. The solid blue line represents the grouped calibration result of the model and the black dotted line represents the ideal reference line.

PFS: Progression-free survival; KM: Kaplan-Meier; GBM: gradient boosting model.

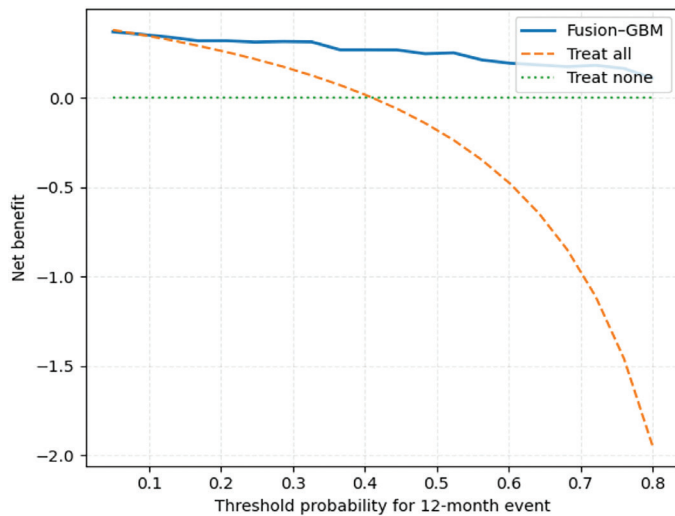


Figure 5. Twelve-month decision curve analysis of the fusion feature set-based gradient boosting model (GBM) on the test set. The solid blue line represents the net benefit of the model, while the orange and green dotted lines correspond to the “treat-all” and “treat-none” strategies.

survival model and found that PET/CT fusion was associated with better predictive performance than single-modality models, but MRI and clinical variables were not incorporated. Another study developed an MRI-based nomogram for the prediction of overall survival, but it remained limited to a single imaging modality [17].

In addition, most machine learning studies in MM have mainly relied on clinical and laboratory variables, with limited integration of multidimensional information such as radiomics, psychological status, and quality of life [18].

In contrast to those studies, the present study attempted to provide a more comprehensive integration of multimodal imaging and clinical information, together with systematic feature construction, model comparison, and performance evaluation. However, the incremental value of this multimodal strategy should be interpreted cautiously because no formal statistical comparison with previously reported models was performed.

Model interpretability is important for the clinical translation of radiomics and machine learning. Existing PET/CT radiomics

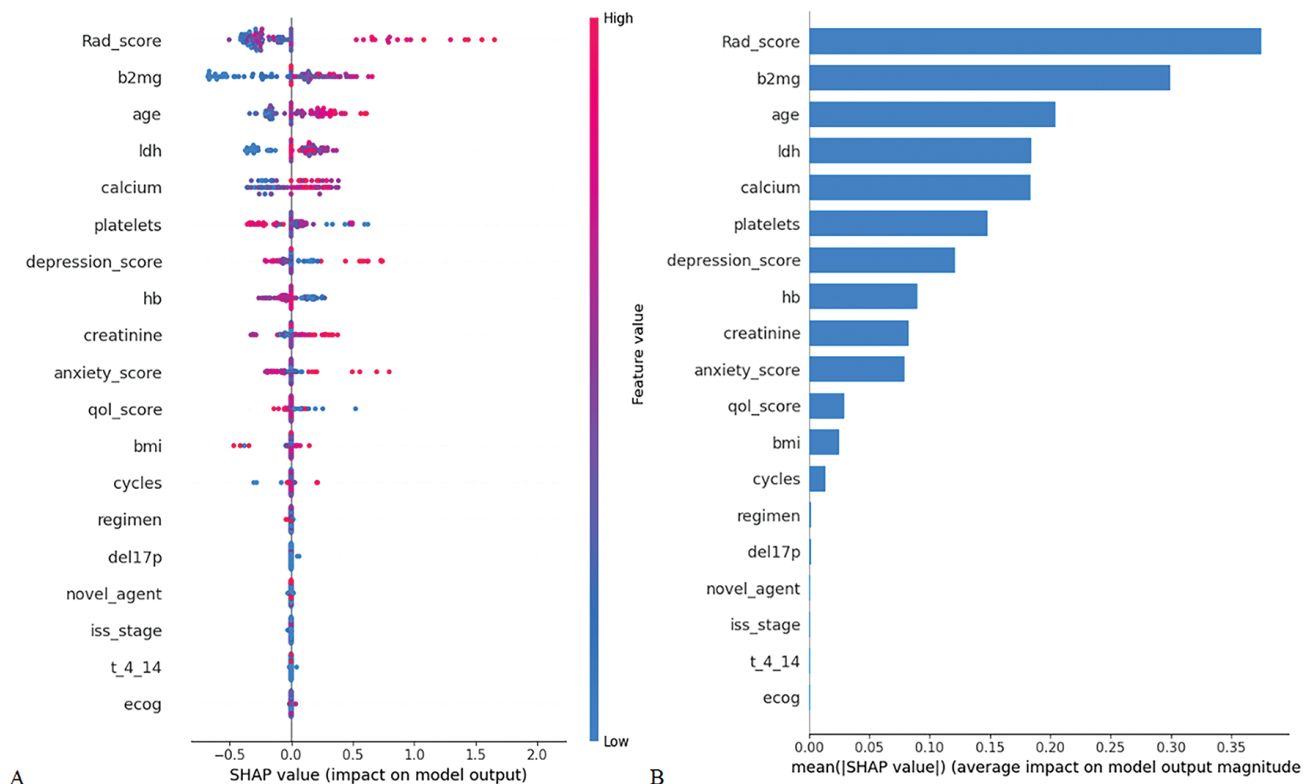


Figure 6. Shapley additive explanations (SHAP) analysis-based interpretation of the fusion feature set-based gradient boosting model (GBM). (A) SHAP summary plot showing the importance and direction of influence of the 24 input features on model prediction. Each dot represents the SHAP value of one patient and the color indicates the feature value. (B) Bar chart of mean SHAP feature importance, with variables ranked by mean absolute SHAP values.

Rad_score: Radiomics-based risk score; b2mg: β 2-microglobulin; ldh: lactate dehydrogenase; hb: hemoglobin; qol: quality of life; bmi: body mass index; iss: International Staging System; ecog: Eastern Cooperative Oncology Group score; del17p: deletion of chromosome 17p; t_4_14: t(4;14) translocation.

models for MM are often limited by black-box characteristics and unstable validation performance [9,19]. In contrast, the present study used SHAP analysis to quantify the contribution of individual features and illustrate how the Rad-score and key clinical indicators may contribute to survival risk estimation.

This approach is consistent with the increasing emphasis on the use of explainable artificial intelligence in hematological malignancies and may help to improve the interpretability of model-based risk prediction [20].

The model also has potential clinical implications. MM shows marked prognostic heterogeneity, and early identification of high-risk patients is crucial for individualized treatment [21]. Existing systems such as the ISS and R-ISS mainly rely on baseline laboratory and cytogenetic variables, including β 2-microglobulin, albumin, LDH, and high-risk cytogenetic abnormalities. Although these systems are widely used, they may not fully reflect tumor spatial heterogeneity, metabolic activity, focal lesion characteristics, or whole-body disease burden captured by multimodal imaging. PET/CT- or MRI-based radiomics models have been explored for MM prognosis, but most previous studies focused on single imaging modalities or did not specifically address early progression risk within a clinically relevant time window [19,22]. Therefore, the proposed fusion model should be considered a potential complement to, rather than a replacement for, the ISS or R-ISS. By focusing on 12-month PFS prediction, the fusion GBM may help to identify patients at increased risk of early progression and support more individualized follow-up, including closer laboratory monitoring, earlier imaging reassessment, and careful evaluation of treatment response. For selected high-risk patients, the model may also provide additional information when considering risk-adapted treatment strategies, such as intensified induction therapy, earlier transplantation, or novel antibody-based agents, together with established staging systems and clinical judgment. Nevertheless, because this study lacked external validation and did not prospectively evaluate whether the model changes decision-making compared to the ISS or R-ISS, its clinical applicability remains preliminary. Future studies should examine whether adding this model to existing risk systems can improve risk classification, treatment decisions, and patient outcomes.

Study Limitations

The present study has some limitations to be noted. First, this was a single-center retrospective study with a relatively small sample size and no independent external validation cohort, which may limit the generalizability of the proposed model. This is particularly relevant for radiomics-based machine learning models, as radiomics features may be affected by differences in imaging protocols, reconstruction settings, preprocessing procedures, and lesion segmentation across institutions. Although standardized preprocessing and manual delineation

were performed by experienced radiologists, potential protocol-related and segmentation-related variability could not be completely avoided. Second, treatment regimens were heterogeneous and residual confounding may remain despite the inclusion of key treatment-related variables in the model. Third, although ISS and R-ISS information was included in the baseline assessment, direct predictive performance comparison or incremental-value analysis against the ISS, R-ISS, or Second Revision of the ISS (R2-ISS) was not performed. Therefore, the proposed model should be regarded as a potential complement to, rather than a replacement for, established prognostic systems. In addition, the R2-ISS score could not be fully evaluated because complete data required for its calculation were unavailable in this retrospective cohort. Future multicenter prospective studies with external validation are needed to directly compare this model with established prognostic systems and further assess its robustness, generalizability, and clinical applicability. Integration of multi-omics data, including genomics, may also improve model performance.

Conclusion

In the present study, various machine learning-based survival models were constructed based on radiomics and clinical features, among which the fusion feature set-based GBM showed numerically favorable performance in the prediction of 12-month PFS, with relatively good discriminative ability, calibration performance, and potential clinical net benefits. The model's transparency and interpretability were further supported by SHAP analysis. The integration of multimodal imaging data and clinical information may contribute to early risk prediction in MM and provide a potential reference for individualized treatment planning after further validation.

Ethics

Ethics Committee Approval: The study was approved by the Ethics Committee of the Affiliated Hospital Six of Nantong University, Yancheng Third People's Hospital (no: YTP2022-026, approved on May 11, 2022).

Informed Consent: Written informed consent was obtained from all patients.

Footnotes

Authorship Contributions

Surgical and Medical Practices: S.S., W.G.; Concept: S.S., Z.H.; Design: S.S., Z.H.; Data Collection or Processing: Y.T., W.G., F.S.; Analysis or Interpretation: S.S., Z.H., Y.T., L.W.; Literature Search: L.W.; Writing: S.S., F.S.

Conflict of Interest: No conflict of interest was declared by the authors.

Financial Disclosure: This study was financially supported by the 2024 Yancheng Science and Technology Bureau Project (no. YCBE202457) and 2024 Research Project of the Yancheng Municipal Health Commission (no. YK2024162).

References

1. Fotiou D, Katodritou E. From biology to clinical practice: the bone marrow microenvironment in multiple myeloma. *J Clin Med*. 2025;14:327.
2. Davies FE, Pawlyn C, Usmani SZ, San-Miguel JF, Einsele H, Boyle EM, Corre J, Auclair D, Cho HJ, Lonial S, Sonneveld P, Stewart AK, Bergsagel PL, Kaiser MF, Weisel K, Keats JJ, Mikhael JR, Morgan KE, Ghobrial IM, Orlowski RZ, Landgren CO, Gay F, Caers J, Chng WJ, Chari A, Walker BA, Kumar SK, Costa LJ, Anderson KC, Morgan GJ. Perspectives on the risk-stratified treatment of multiple myeloma. *Blood Cancer Discov*. 2022;3:273-284.
3. Rajkumar SV. Multiple myeloma: 2022 update on diagnosis, risk stratification, and management. *Am J Hematol*. 2022;97:1086-1107.
4. Spinnato P, Filonzi G, Conficoni A, Facchini G, Ponti F, Sambri A, De Paolis M, Cavo M, Salizzoni E, Nanni C. Skeletal survey in multiple myeloma: role of imaging. *Curr Med Imaging*. 2021;17:956-965.
5. Wu F, Bernard S, Fayad LM, Ilaslan H, Messiou C, Mouloupoulos LA, Mulligan ME. Updates and ongoing challenges in imaging of multiple myeloma: *AJR* expert panel narrative review. *AJR Am J Roentgenol*. 2021;217:775-785.
6. Michalska-Foryszewska A, Rogowska A, Kwiatkowska-Miernik A, Sklinda K, Mruk B, Hus I, Walecki J. Role of imaging in multiple myeloma: a potential opportunity for quantitative imaging and radiomics? *Cancers*. 2024;16:4099.
7. Chen H, Han J, Huang H, He Q, Ren X, Yu F, Chang C, Ding X, Luo Q. A feasibility study of [18F] FDG PET/CT radiomics in predicting high-risk cytogenetic abnormalities in multiple myeloma. *EJNMMI Res*. 2025;15:131.
8. Sachpekidis C, Goldschmidt H, Edenbrandt L, Dimitrakopoulou-Strauss A. Radiomics and artificial intelligence landscape for [18F] FDG PET/CT in multiple myeloma. *Semin Nucl Med*. 2025;55:387-395.
9. Klontzas ME, Triantafyllou M, Leventis D, Koltsakis E, Kalarakis G, Tzortzakakis A, Karantanias AH. Radiomics analysis for multiple myeloma: a systematic review with radiomics quality scoring. *Diagnostics (Basel)*. 2023;13:2021.
10. Romero M, Mosquera Orgueira A, Mejía Saldarriaga M. How artificial intelligence revolutionizes the world of multiple myeloma. *Front Hematol*. 2024;3:1331109.
11. Ren L, Xu B, Xu J, Li J, Jiang J, Ren Y, Liu P. A machine learning model to predict survival and therapeutic responses in multiple myeloma. *Int J Mol Sci*. 2023;24:6683.
12. Rajkumar SV, Dimopoulos MA, Palumbo A, Blade J, Merlini G, Mateos MV, Kumar S, Hillengass J, Kastritis E, Richardson P, Landgren O, Paiva B, Dispenzieri A, Weiss B, LeLeu X, Zweegman S, Lonial S, Rosinol L, Zamagni E, Jagannath S, Sezer O, Kristinsson SY, Caers J, Usmani SZ, Lahuerta JJ, Johnsen HE, Beksac M, Cavo M, Goldschmidt H, Terpos E, Kyle RA, Anderson KC, Durie BG, San Miguel JF. International Myeloma Working Group updated criteria for the diagnosis of multiple myeloma. *Lancet Oncol*. 2014;15:e538-e548.
13. Landgren O, Prior TJ, Masterson T, Heuck C, Bueno OF, Dash AB, Einsele H, Goldschmidt H, Knop S, Li C, Mellqvist UH, McFadden I, Oprea C, Ross JA, Talpes M, Hydren JR, Ahlstrom JM, Kazandjian D, Weinhold N, Zhang R, Stetler-Stevenson M, Marti G, Devlin SM. EVIDENCE meta-analysis: evaluating minimal residual disease as an intermediate clinical end point for multiple myeloma. *Blood*. 2024;144:359-367.
14. Zhang T, Lin Z, Zheng Z, Wang Q, Zhou S, Zhang B, Zheng D, Chen Z, Zheng S, Zhang Y, Lin X, Dong R, Chen J, Qian H, Hu X, Zhuang Y, Zhang Q, Jin Z, Jiang S, Ma Y. Prognostic significance of $\beta 2$ -microglobulin decline index in multiple myeloma. *Front Oncol*. 2024;14:1322680.
15. Du C, Li L, Fan H, Mao X, Liu J, Xu Y, Sui W, Deng S, Li C, Zhao J, Yi S, Hao M, Zou D, Zhao Y, Qiu L, An G. The age-dependent changes in risk weights of the prognostic factors for multiple myeloma. *Hematology*. 2023;28:2258686.
16. Guinea-Pérez J, Ceresi A, Fernández García A, Arroyo Galende B, Belmonte-Hernández A, Peluso S, Alvarez F. Radiomics feature analysis for survival prediction in multiple myeloma: an automated PET/CT approach. *Comput Methods Programs Biomed*. 2025;271:109019.
17. Li Y, Liu Y, Yin P, Hao C, Sun C, Chen L, Wang S, Hong N. MRI-based bone marrow radiomics nomogram for prediction of overall survival in patients with multiple myeloma. *Front Oncol*. 2021;11:709813.
18. Sharma R, Desai K, Dadhanian N, Croce P, Thalji M, Hanif A, Mirza W, ElManzalawi Y. Systematic review of currently available predictive machine learning models for multiple myeloma. *Blood*. 2024;144(Suppl 1):7518.
19. Ni B, Huang G, Huang H, Wang T, Han X, Shen L, Chen Y, Hou J. Machine learning model based on optimized radiomics feature from 18F-FDG-PET/CT and clinical characteristics predicts prognosis of multiple myeloma: a preliminary study. *J Clin Med*. 2023;12:2280.
20. Nazha A, Elemento O, Ahuja S, Lam B, Miles M, Shouval R, McWeeney S, Sirhan S, Srisuwananukorn A, Haferlach T. Artificial intelligence in hematology. *Blood*. 2025;146:2283-2292.
21. Schavgoulidze A, Perrot A, Leleu X, Cazaubiel T, Chretien ML, Feugier P, Belhadj K, Manier S, Roussel M, Brechignac S, Orsini-Piocelle F, Mohty M, Schiano de Collella JM, Macro M, Adiko D, Dib M, Fontan J, Luttiau-Motard C, Bouscary D, Pascal L, Roland V, Lifermann F, Bakala J, Montes L, Kennel C, Rey P, Richez V, Keddar F, Frenzel L, Calmettes C, Chaleteix C, Plantier I, Chalayer E, Schmitt A, Roul C, Demarquette H, Cerutti C, Pavageau L, Derrier L, Avet-Loiseau H, Corre J. High-risk genomic consensus validation for patients with newly diagnosed multiple myeloma using next-generation sequencing. *Blood*. 2026;147:266-275.
22. Yu T, Chen M, Huang B, Liu J, Li X, Gu J, Li Y, Li J. Clinical feature and predictive model for transplanted patients with functional high-risk multiple myeloma. *Arch Med Res*. 2026;57:103291.