BMC Cancer

RESEARCH

Open Access

Association of sarcopenia with all-cause and cause-specific mortality in cancer patients: development and validation of a 3-year and 5-year survival prediction model

Feng Cui[1†], Xiangji Dang[2†], Daiyun Peng[3], Yuanhua She[1], Yubin Wang[4], Ruifeng Yang[5], Zhiyao Han[5], Yan Liu[6*] and Hanteng Yang[1*]

Abstract

Background Sarcopenia is a clinicopathological condition characterized by a decrease in muscle strength and muscle mass, playing a crucial role in the prognosis of cancer. Therefore, this study aims to investigate the association between sarcopenia and both all-cause mortality and cancer-specific mortality among cancer patients. Furthermore, we plan to develop risk prediction models using machine learning algorithms to predict 3-year and 5-year survival rates in cancer patients.

Method This study included 1095 cancer patients from the National Health and Nutrition Examination Survey (NHANES) cohorts spanning 1999–2006 and 2011–2014. Initially, we used the Least Absolute Shrinkage and Selection Operator (LASSO)-Cox regression models for feature selection. Subsequently, we employed multivariable Cox regression models to investigate the association between sarcopenia and all-cause and cancer-specific mortality in cancer patients. We developed five machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), LightGBM, and XGBoost, to predict 3-year and 5-year survival rates and to perform risk stratification.

Results The multivariable COX regression model showed sarcopenia significantly increases the risk of all-cause mortality (HR = 1.33, 95%CI:1.05, 1.70, $P = 0.0194$) and cancer-specific mortality (HR = 1.67, 95%CI:1.09, 2.55, $P = 0.0176$) in cancer patients. Among the five machine learning algorithms developed, the LightGBM model demonstrated strong performance in the 3-year and 5-year survival prediction tasks, making it the optimal model selection. Decision curve analysis and Kaplan–Meier curves further confirmed our model's ability to identify high-risk individuals effectively.

Conclusions Sarcopenia significantly increases the risk of mortality in cancer patients. We developed a survival prediction model for cancer patients that effectively identifies high-risk individuals, thereby providing a foundation for personalized survival assessment.

Keywords Sarcopenia, Cancer, Mortality, Machine Learning, Risk assessment

[†]Feng Cui and Xiangji Dang contributed equally to this work.

*Correspondence:
Yan Liu
liuyanly@lzu.edu.cn
Hanteng Yang
yanghanteng@126.com
Full list of author information is available at the end of the article

Introduction

Sarcopenia is a clinicopathological condition characterized by a reduction in muscle strength and muscle mass [1, 2], which is more prevalent in the aged and cancer patients [3]. Current research indicates that sarcopenia is closely associated with a variety of adverse health outcomes, including aging [4], obesity [5], chronic inflammation [6], cardiovascular diseases [7], malignancies [8], and an increased risk of mortality [9].

Cancer, as a significant global public health issue, has become a leading cause of death, with an estimated 9.7 million cancer-related deaths worldwide in 2022 [10]. In recent years, as an understanding of cancer management has deepened, more studies have found that nutritional status, muscle condition, and weight changes play critical roles in the development and prognosis of cancer [11]. For cancer patients, sarcopenia not only affects their body composition but is also associated with various adverse clinical outcomes, including prolonged hospitalization, increased postoperative complications, reduced chemotherapy tolerance, and shortened survival [12]. Therefore, in-depth research on the impact of sarcopenia on the prognosis of cancer patients has significant clinical importance. Although existing studies have shown that sarcopenia is related to both short-term and long-term outcomes in cancer patients [13–15], the relationship between sarcopenia and mortality in cancer patients has not been fully explored.

With the rapid advancement of artificial intelligence, machine learning (ML) has emerged as an efficient and personalized solution for addressing complex medical problems by automatically identifying patterns and features within data. Compared with traditional statistical methods, ML offers superior predictive power and generalizability, and has thus been widely applied in various areas of medical research, including disease diagnosis, prognosis evaluation, and individualized treatment decision-making—particularly in the field of oncology [16, 17]. For example, an ML model based on radiomics demonstrated favorable performance in predicting the response of locally advanced gastric cancer to neoadjuvant immunotherapy [18]. In addition, ML models constructed using clinicopathological features have shown satisfactory accuracy in predicting survival outcomes in patients with residual gastric cancer [19]. By leveraging such techniques, multidimensional clinical data can be effectively utilized to provide more accurate and personalized survival predictions for cancer patients.

Therefore, this study aims to investigate the relationship between sarcopenia and mortality among cancer patients based on the NHANES database. Furthermore, we strive to develop a risk prediction model using machine learning algorithms to predict cancer patients' 3-year and 5-year survival rates. This model will help identify high-risk individuals and provide a basis for early prevention and personalized treatment.

Methods

Study population

NHANES is a cross-sectional survey conducted by the National Center for Health Statistics (NCHS). It is designed using stratified, multistage probability sampling to provide a reliable assessment of the nutritional and health status of the population in the U.S. [20]. The NCHS Ethics Review Board approved the NHANES study, and all participants provided written informed consent.

The study's data was obtained from six cycles of NHANES, spanning from 1999 to 2006 and 2011 to 2014. These cycles include accessible whole-body dual-energy X-ray absorptiometry (DXA) data, encompassing a total of 61,405 participants. Subsequently, the information of these participants was linked to the National Death Index (NDI) from the NCHS, thereby constructing the NHANES longitudinal follow-up cohort [21]. Follow-up was conducted until December 31, 2019, to determine the survival status of these individuals. Our study recruited adult participants eligible for mortality follow-up, excluding those under 20 years of age, not available for public release, or ineligible for mortality follow-up.

Furthermore, individuals without height and DXA measurements, as well as those who did not disclose their cancer or malignancy status on the questionnaire or had missing questionnaire information, were excluded. Cancer or malignancy diagnoses are based on a question asked by trained interviewers: 'Have you ever been told by a doctor or other health professional that you had cancer or a malignancy of any kind?' The study included 1095 subjects. The process of data selection is detailed in Fig. 1. All data used in this study are publicly available.

Sarcopenia and Mortality

The definition of sarcopenia is based on the revised consensus of the European Working Group on Sarcopenia in Older People (EWGSOP) [1]. Sarcopenia is diagnosed when muscle mass falls below 7.0 kg/m^2 for men and 5.5 kg/m^2 for women. Muscle mass is measured using appendicular skeletal muscle mass (ASM), which is obtained by adjusting the square of height. Dual-energy X-ray absorptiometry (DXA) is a non-invasive method widely used today to measure muscle mass [22]. Muscle mass is the sum of appendicular skeletal muscle mass measured by DXA. DXA measurements were performed using a Hologic QDR 4500A fan-beam X-ray bone densitometer (Hologic Inc., Marlborough, MA, USA) and a Hologic Discovery A densitometer in NHANES

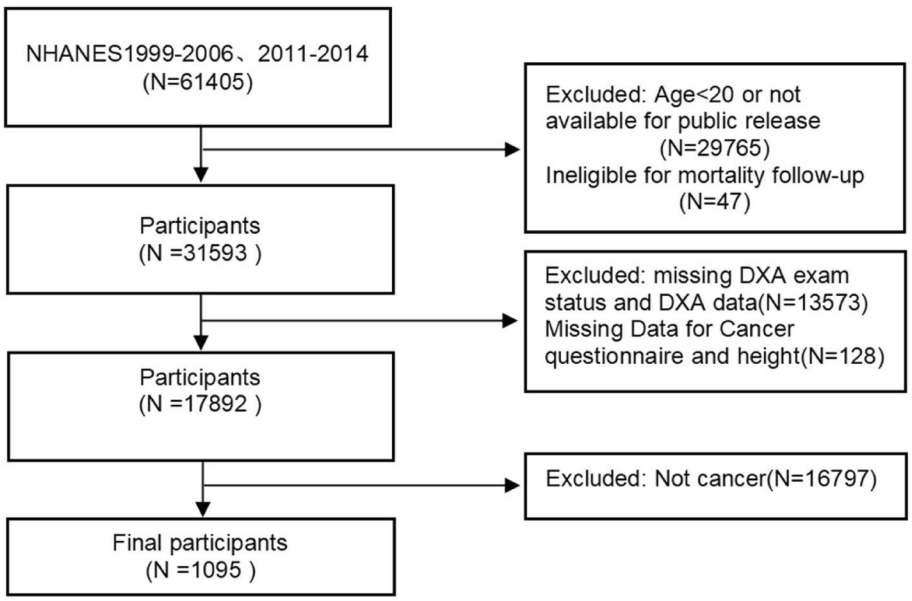

Fig. 1 Flowchart for participant selection in this study

1999–2006 and 2011–2014, respectively. Height was measured at the mobile screening center using standardized techniques and equipment. Pregnant women weighing over 136 kg or measuring over 196 cm tall were excluded from DXA measurements to ensure safety. All-cause mortality was determined based on records from the NDI. Specific mortality was determined using ICD-10. Cancer deaths were defined as ICD-10 codes C00-C97.

Covariates

Based on current research regarding sarcopenia and cancer, this study included additional variables potentially associated with both sarcopenia and mortality among cancer patients as covariates. This study incorporated 21 covariates, including demographic characteristics, socioeconomic status, lifestyle factors, medical history, physical measurement indicators, and laboratory test markers. Demographic characteristics encompassed gender, age, ethnicity, and marital status. The poverty income ratio and educational level represented socioeconomic indicators. Lifestyle factors primarily involved smoking and alcohol consumption habits. Medical history included diabetes, hypertension, and coronary heart disease. Physical measurements comprised body mass index (BMI). Laboratory tests covered ten variables: white blood cell count, hemoglobin, albumin, and creatinine. The data for BMI and alcohol intake were collected by trained health technicians at Mobile Examination Centers (MEC). The remaining variables were gathered through standardized questionnaires during household interviews. Further details on definitions can be found in the Additional file 1: Supplementary methods. All variables had missing rates < 10%, and we filled in the missing values using the missForest method, which interpolates non-parametric missing values for mixed-type data [23].

Feature selection

Univariate Cox regression analysis examined the associations between variables and survival time. Subsequently. LASSO-Cox regression was utilized to mitigate the effects of multicollinearity among variables and to generate the final feature set. This technique identifies irrelevant features by zeroing their coefficients, dependent on the regularization parameter λ. The optimal λ was ascertained via tenfold cross-validation, choosing the value that minimized the mean standard error.

Multivariable COX regression model

Before constructing the Cox proportional hazards model, we assessed potential violations of the proportional hazards assumption by examining whether Kaplan–Meier curves for all-cause and cancer-specific mortality crossed according to sarcopenia status. Subsequently, Multivariable Cox regression models were employed to investigate the relationship between sarcopenia and the risks of all-cause and cancer-specific mortality in cancer patients. Hazard ratios (HRs) and their 95% confidence intervals (CIs) were estimated. Additionally, the association between sarcopenia and cancer-specific mortality was further evaluated using the Fine-Gray subdistribution hazard model, considering non-cancer death as a

competing event. Subdistribution hazard ratios (SHRs) and their 95% CIs were calculated accordingly.

Furthermore, this study investigated potential interactions between multiple stratification variables and mortality, and sensitivity analyses were conducted to verify the robustness of the findings. To ensure the accuracy of the analyses, patients who died within the first two years of the follow-up period were excluded, thus further validating the association between sarcopenia and long-term mortality in cancer patients.

Machine learning models and risk stratification

To address sarcopenia, we developed five machine learning algorithms, including Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), LightGBM, and XGBoost, to predict the 3-year and 5-year survival rates of patients. The dataset was randomly split into training and validation sets in an 8:2 ratio. To ensure model stability and generalizability, we employed a fivefold cross-validation approach. The predictive performance of each model was evaluated by drawing ROC curves and calculating multiple performance metrics such as accuracy, AUC, sensitivity, and specificity to identify the model with the best average performance as the final predictive tool. Additionally, the clinical utility of the final model was explored through Decision Curve Analysis (DCA). Based on the risk scores from the final model, optimal risk thresholds were established, categorizing patients into high-risk and low-risk groups. Kaplan–Meier survival curves illustrated significant survival differences between these risk groups, and the Log-rank test was used to verify these differences statistically.

Statistical analyses

Baseline characteristics are reported as medians and interquartile ranges for continuous variables, compared with Mann–Whitney U test, and as frequencies and percentages for categorical variables, analyzed with Chi-squared tests. The median follow-up time was estimated using the reverse Kaplan–Meier method. Data analysis and visualization were performed using R software version 4.1 (available at http://www.R-project.org; provided by The R Foundation), EmpowerStats (available at http://www.empowerstats.com; provided by X&Y Solutions, Inc.), and Python version 3.7.12 (available at https://www.python.org). Machine learning algorithms were implemented using scikit-learn version 1.0.2. The optimal cut-off value for risk scoring was determined using X-tile software version 3.6.1 [24]. A two-sided p-value of less than 0.05 was considered statistically significant.

Results

Baseline characteristics of study participants

This study ultimately included 1095 cancer patients who met the follow-up criteria, with females comprising 55.71%. During a median follow-up of 197 months, 471 deaths were recorded, including 155 cancer-specific deaths. Table 1 presents the baseline characteristics of the patients, categorized by the presence or absence of sarcopenia. Patients with sarcopenia were more likely to be older, non-Hispanic white, and current drinkers (all $P<0.05$). Baseline characteristics according to all-cause mortality are detailed in Additional file 1: Supplementary Table 1. Baseline characteristics of patients with cancer types having a sample size greater than 50, stratified by sarcopenia status, are presented in Additional File 1: Supplementary Table 2.

Feature selection

The results of univariate Cox regression analyses for all variables are presented in Additional file 1: Supplementary Table 3, with covariates exhibiting P-values less than 0.05 subsequently included in further LASSO-Cox selection. Through tenfold cross-validation to maximize the Concordance index, we identified the optimal λ value that yielded the highest Concordance index (Fig. 2B). Using this optimal λ value, we fitted the LASSO-Cox regression model, identified features with coefficients equal to zero, and subsequently removed them(Fig. 2A). This process resulted in retaining 17 features with non-zero coefficients (Fig. 2C).

Figure 2A and B depict the coefficients and concordance indices from tenfold cross-validation. Figure 2C Illustrates the selected final feature set.

Association between sarcopenia and mortality

Kaplan–Meier curves indicated that cancer patients with sarcopenia had significantly higher all-cause and cancer-specific mortality rates compared to those without sarcopenia (Log-rank $P<0.001$), suggesting no violation of the proportional hazards assumption of the Cox model (Fig. 3).

Multivariable Cox regression analysis indicated that sarcopenia is significantly associated with higher all-cause and cancer-specific mortality in cancer patients. Table 2 illustrates that in the unadjusted crude model, HRs for all-cause and cancer-specific mortality for patients with sarcopenia were 2.06 (95% CI: 1.70–2.50, $P<0.0001$) and 1.97 (95% CI: 1.41- 2.76, $P<0.0001$), respectively. After adjusting for age and gender, the HR for all-cause mortality decreased to 1.77 (95% CI: 1.45–2.15, $P<0.0001$), and the HR for cancer-specific mortality reduced to 1.79 (95% CI: 1.27–2.52, $P=0.0008$).

Table 1 Baseline characteristics of participants stratified by sarcopenia status

	Total	Non-sarcopenia	Sarcopenia	*P*-value
PIR	2.73 (1.45, 4.65)	2.93 (1.57, 4.99)	2.16 (1.30, 3.75)	< 0.001
BMI, kg/m2	26.90 (23.90, 30.57)	28.08 (25.38, 31.79)	22.80 (20.77, 24.38)	< 0.001
ALT, IU/L	20.00 (16.00, 26.00)	21.00 (17.00, 27.00)	18.00 (15.00, 22.00)	< 0.001
Albumin, g/dl	4.25 (4.10, 4.40)	4.20 (4.00, 4.40)	4.30 (4.10, 4.50)	0.031
Creatinine, μmol/L	79.56 (61.90, 88.40)	79.56 (61.90, 88.40)	79.56 (61.90, 97.20)	0.684
Uric acid, mg/dl	5.30 (4.40, 6.10)	5.40 (4.60, 6.30)	4.80 (3.92, 5.70)	< 0.001
Blood urea nitrogen, mmol/L	5.00 (3.93, 6.40)	5.00 (3.93, 6.10)	5.04 (3.93, 6.43)	0.420
Total cholesterol, mmol/L	5.17 (4.53, 5.82)	5.20 (4.53, 5.87)	5.08 (4.54, 5.67)	0.188
Red blood cell count, 10^6 cells/μL	4.55 (4.29, 4.89)	4.57 (4.30, 4.90)	4.50 (4.26, 4.83)	0.032
Hemoglobin, g/dl	14.10 (13.30, 15.00)	14.10 (13.30, 15.00)	14.00 (13.40, 14.90)	0.407
White blood cell count, 10^3cells/μL	6.80 (5.60, 8.10)	6.80 (5.60, 8.06)	6.79 (5.60, 8.20)	0.629
Age, years				< 0.001
< 65	593 (54.16%)	507 (58.61%)	86 (37.39%)	
≥ 65	502 (45.84%)	358 (41.39%)	144 (62.61%)	
Sex, *n* (%)				0.751
Female	610 (55.71%)	484 (55.95%)	126 (54.78%)	
Male	485 (44.29%)	381 (44.05%)	104 (45.22%)	
Race, *n* (%)				< 0.001
Hispanic	126 (11.51%)	102 (11.79%)	24 (10.43%)	
Non-Hispanic Black	132 (12.05%)	122 (14.10%)	10 (4.35%)	
Non-Hispanic White	802 (73.24%)	614 (70.98%)	188 (81.74%)	
Other Race	35 (3.20%)	27 (3.12%)	8 (3.48%)	
Marital, *n* (%)				0.567
Never married	59 (5.48%)	46 (5.39%)	13 (5.86%)	
Married /Living with partner	697 (64.78%)	561 (65.69%)	136 (61.26%)	
Widowed/Divorced/ Separated	320 (29.74%)	247 (28.92%)	73 (32.88%)	
Education, *n* (%)				0.082
< high school	249 (22.78%)	185 (21.41%)	64 (27.95%)	
high school	266 (24.34%)	217 (25.12%)	49 (21.40%)	
> high school	578 (52.88%)	462 (53.47%)	116 (50.66%)	
Smoking, *n* (%)				0.584
never	456 (41.64%)	367 (42.43%)	89 (38.70%)	
now	217 (19.82%)	170 (19.65%)	47 (20.43%)	
former	422 (38.54%)	328 (37.92%)	94 (40.87%)	
Drinking, *n* (%)				0.025
never	138 (12.96%)	102 (12.14%)	36 (16.00%)	
now	651 (61.13%)	531 (63.21%)	120 (53.33%)	
former	276 (25.92%)	207 (24.64%)	69 (30.67%)	
Diabetes, *n* (%)				0.219
no	893 (81.55%)	699 (80.81%)	194 (84.35%)	
yes	202 (18.45%)	166 (19.19%)	36 (15.65%)	
Hypertension, *n* (%)				0.595
no	464 (42.37%)	363 (41.97%)	101 (43.91%)	
yes	631 (57.63%)	502 (58.03%)	129 (56.09%)	
Coronary heart disease, *n* (%)				0.552
no	1004 (92.19%)	795 (92.44%)	209 (91.27%)	
yes	85 (7.81%)	65 (7.56%)	20 (8.73%)	
Status, *n* (%)				< 0.001
Alive	606 (55.34%)	524 (60.58%)	82 (35.65%)	
Death	489 (44.66%)	341 (39.42%)	148 (64.35%)	

Median (Q1, Q3) for continuous variables, *n* (%) for categorical variables

BMI Body mass index, *PIR* Poverty income ratio

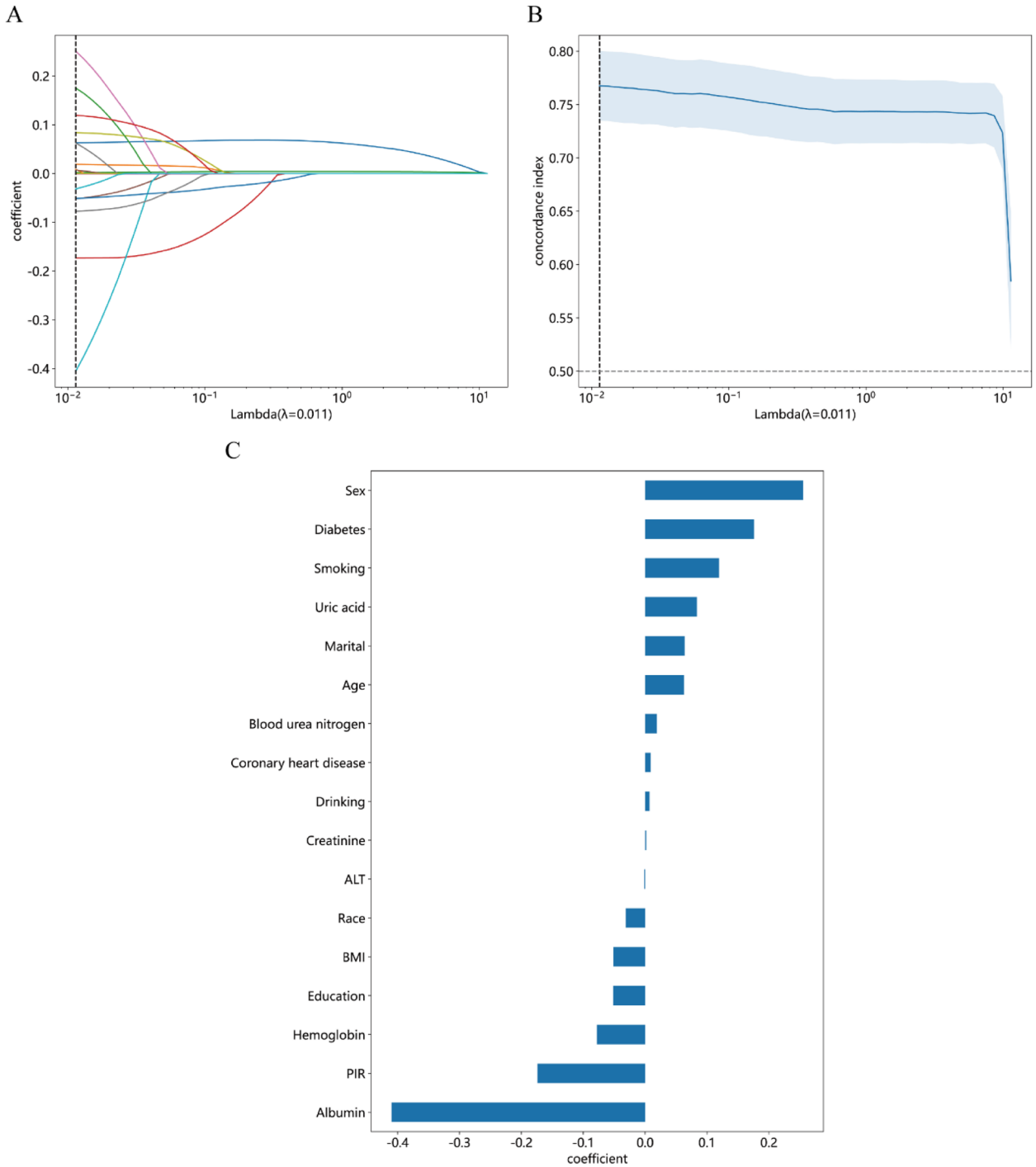

Fig. 2 The feature selection process uses the LASSO-Cox regression model

Further adjustments in Model 2 led to a further decrease in the HR for all-cause mortality to 1.30 (95% CI: 1.02–1.64, $P=0.0311$), which remained significant; however, the HR for cancer-specific mortality changed slightly. After total adjustment, Model 3 showed that sarcopenia was associated with an HR of 1.33 (95% CI: 1.05–1.70; $P=0.0149$) for all-cause mortality and an HR of 1.67 (95% CI: 1.09–2.55; $P=0.0176$) for cancer-specific mortality. Furthermore, the association between sarcopenia and cancer-specific mortality remained robust in the Fine and Gray competing risk model (Additional file 1: Supplementary Table 4).

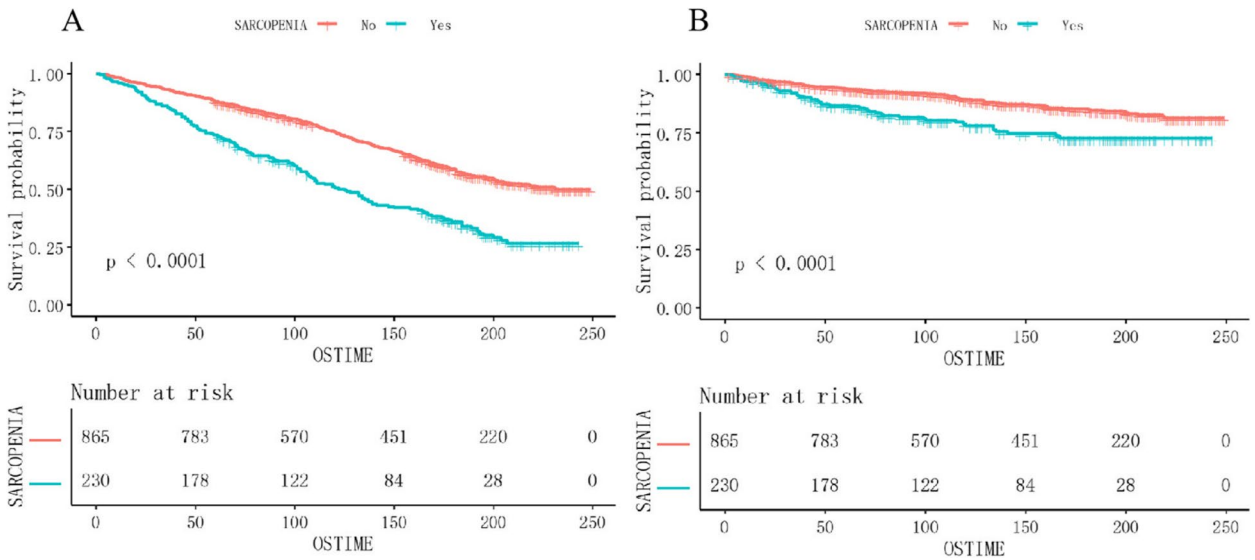

Fig. 3 Kaplan–Meier curves for different mortality outcomes in the total population. Kaplan–Meier survival curves for A all-cause mortality and B cancer mortality among total participants grouped by sarcopenia status

Table 2 Relationship between sarcopenia and mortality in cancer patients

Sarcopenia	Crude Model HR (95%CI) *P*-value	Model 1 HR (95%CI) *P*-value	Model 2 HR (95%CI) *P*-value	Model 3 HR (95%CI) *P*-value
All-cause mortality				
No	1.0(reference)	1.0(reference)	1.0(reference)	1.0(reference)
Yes	2.06 (1.70, 2.50) < 0.0001	1.77 (1.45, 2.15) < 0.0001	1.30 (1.02, 1.64) 0.0311	1.33 (1.05, 1.70) 0.0194
Cancer mortality				
No	1.0(reference)	1.0(reference)	1.0(reference)	1.0(reference)
Yes	1.97 (1.41, 2.76) < 0.0001	1.79 (1.27, 2.52) 0.0008	1.63 (1.07, 2.47) 0.0228	1.67 (1.09, 2.55) 0.0176

Data are presented as HR (95%CI)

Model 1: adjust for Age, Sex

Model 2: adjust for Age; Sex; Race; Marital; PIR; Education; Smoking; Drinking; Diabetes; Coronary heart disease; BMI

Model 3: adjust for Age; Sex; Race; Marital; PIR; Education; Smoking; Drinking; Diabetes; Coronary heart disease; BMI; ALT; Albumin; Creatinine; Uric acid; Blood urea nitrogen; Hemoglobin

Subgroup analyses and sensitivity analyses

In this study, we also conducted subgroup analyses and interaction tests (Fig. 4) to explore how other potential influencing factors, such as age, gender, and race, interact with the independent effects of sarcopenia on mortality risk. Subgroup analysis revealed that sarcopenia increased the risk of both all-cause and cancer-specific mortality in both men and women, although the risk increase was more pronounced in men. However, the interaction between gender and sarcopenia was not statistically significant in the adjusted models ($P > 0.05$). Notably, sarcopenia was associated with a 452% increase in all-cause mortality risk and a 939% increase in cancer-specific mortality among non-Hispanic Black individuals. Interaction analysis indicated that, except for race and alcohol consumption, most variables, such as education level, diabetes status, and body mass index (BMI), did not reach statistical significance in their interactions with sarcopenia ($P > 0.05$), suggesting the reliability of our study's findings. Interestingly, we observed a downward trend in the risk of all-cause and cancer-specific mortality as BMI increased. The association between sarcopenia and mortality varied across different cancer subtypes (Additional file 1: Supplementary Table 5), with stronger associations observed in solid tumors. However, this finding should be interpreted with caution due to the limited sample sizes. Sensitivity analyses confirmed the consistency of our results even after excluding cancer patients who died within the first two years of follow-up (Additional file 1: Supplementary Table 6), further validating the stability and reliability of our findings.

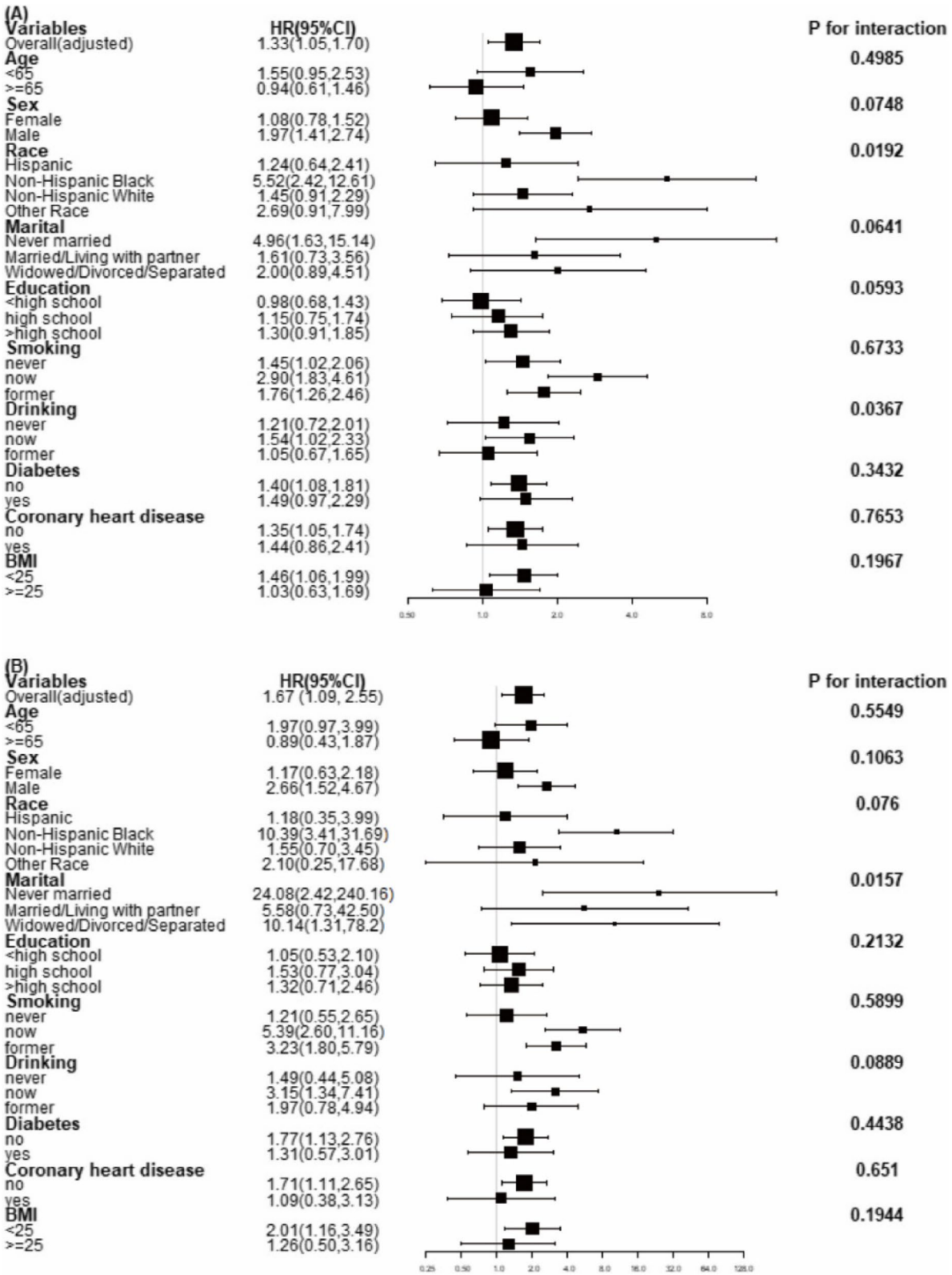

Fig. 4 Forest plot of subgroup analyses of the relationship between sarcopenia and mortality. Forest plots showing the association of sarcopenia with **A** all-cause mortality and **B** cancer mortality among total participants. adjust for Age; Sex; Race; Marital; PIR; Education; Smoking; Drinking; Diabetes; Coronary heart disease; BMI; ALT; Albumin; Creatinine; Uric acid; Blood urea nitrogen; Hemoglobin. HR, hazard ratio; CI, confidence interval

Machine learning

The baseline clinical characteristics of patients within the machine learning dataset are available in Additional file 1: Supplementary Table 7. Substantial differences between the training and test sets were observed for only some features. Table 3 presents the performance metrics of all models used to predict 3-year and 5-year survival rates. Based on a comprehensive evaluation of accuracy, AUC, sensitivity, specificity, and F1 scores, along with the precision reflected by confidence intervals, LightGBM demonstrated superior performance in both tasks. Specifically, LightGBM excelled in the 3-year prediction task with an AUC of 0.842 (95% CI: 0.766–0.919) and an F1 score of 0.830. Although XGBoost exhibited a slightly higher AUC (0.891; 95% CI: 0.840–0.942), LightGBM maintained consistent and robust performance in the 5-year prediction task, characterized by a moderate confidence interval width (AUC = 0.795; 95% CI: 0.723–0.867), indicating acceptable precision and stability. In contrast, other models exhibited either wider confidence intervals or lower overall performance metrics, diminishing their reliability. Based on these results, LightGBM was selected as the optimal predictive model.

The ROC analysis results of each model in the test cohort for predicting 3-year and 5-year survival rates are displayed in Fig. 5. This study also evaluated the LightGBM model through Decision Curve Analysis (DCA). Decision curves for the 3-year and 5-year survival prediction models are displayed in Fig. 6. Compared to scenarios without a prediction model (i.e., treat-all or treat-none scheme), the 3-year and 5-year survival prediction models showed significant benefits for interventions in patients with a prediction probability.

Risk stratification

The 3-year and 5-year survival risk scores for each patient were generated based on the constructed LightGBM prediction model, with the risk score range for 3-year survival being 0.5–1.00 and the optimal cut-off value set at 0.8, and for 5-year survival, the risk score range was 0.4–0.9 with the optimal cut-off value set at 0.8. Based on these cut-off values, individuals were classified into low-risk (3-year risk score $\geq$ 0.8, 5-year risk score $\geq$ 0.8) and high-risk groups (3-year risk score < 0.8, 5-year risk score < 0.8) (Additional file 1: Supplementary Fig. 1). Kaplan–Meier curves demonstrated that the low-risk groups exhibited higher survival rates than high-risk groups, and the differences in survival rates were statistically significant, as indicated by the Log-rank test (*P*-values < 0.001). This suggests that our model can effectively identify high-risk populations (Fig. 7).

Discussion

In this study conducted within the NHANES cohort of cancer respondents, we found that sarcopenia is significantly associated with an increased risk of both all-cause mortality and cancer-specific mortality among cancer survivors. After excluding participants who died within the first two years of follow-up, the results remained consistent, indicating that sarcopenia is linked to poorer long-term prognosis in cancer patients. Subsequently, we developed and validated machine learning models to predict cancer survivors' 3-year and 5-year survival rates based on sarcopenia. These models not only aid in identifying individuals at high risk of mortality and provide a practical foundation for developing personalized prevention and intervention strategies.

Our research demonstrates a significant correlation between sarcopenia and increased risks of both all-cause mortality and cancer-specific mortality among

Table 3 Comparative analysis of performance metrics of various machine learning models

Task	Model_name	Accuracy	AUC	Sensitivity	Specificity	F1
3 year survival	LR	0.749	0.797	0.751	0.722	0.846
	SVM	0.868	0.851	0.876	0.778	0.924
	RandomForest	0.082	0.799	0.000	1.000	NaN
	XGBoost	0.795	0.891	0.781	0.944	0.875
	LightGBM	0.731	0.842	0.716	0.889	0.830
5 year survival	LR	0.799	0.856	0.801	0.788	0.871
	SVM	0.735	0.774	0.731	0.758	0.824
	RandomForest	0.466	0.801	0.376	0.970	0.545
	XGBoost	0.621	0.742	0.581	0.848	0.722
	LightGBM	0.726	0.795	0.715	0.788	0.816

LR Logistic regression, *RF* Random forest, *XGBoost* Extreme gradient boosting, *LightGBM* Light gradient boosting machine, *SVM* Support vector machine, *AUC* Area under the curve

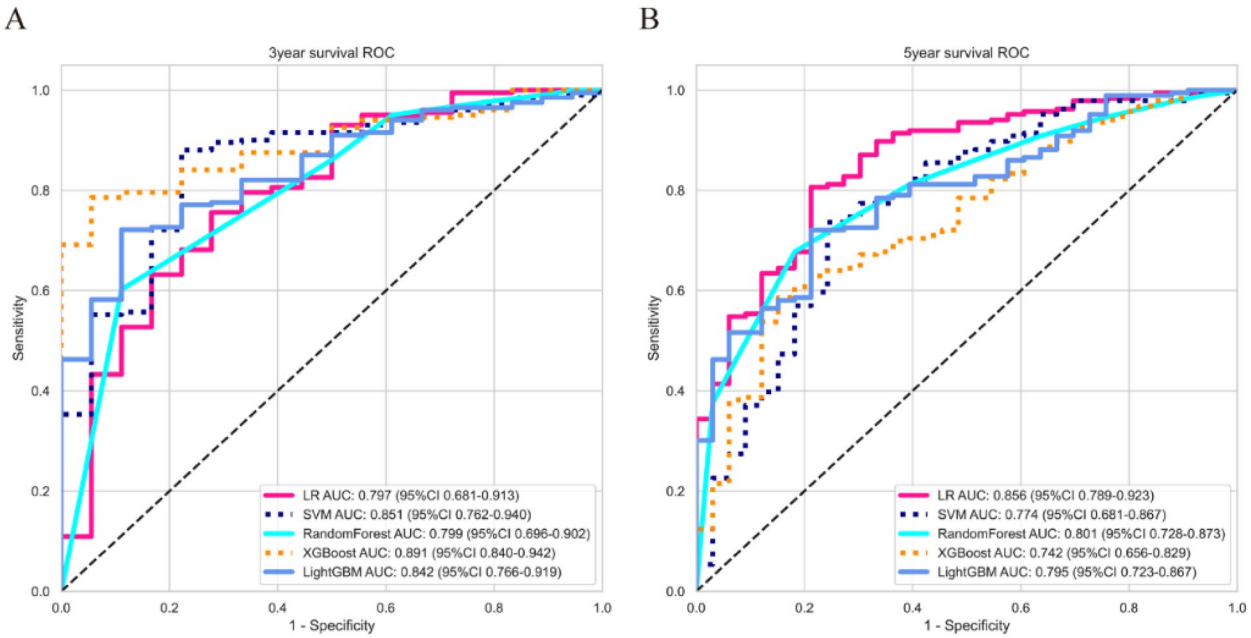

Fig. 5 ROC curve analysis of five machine learning algorithms in the test cohort. **A** displays the ROC curve for the 3-year survival prediction model, while **B** shows the ROC curve for the 5-year survival prediction model

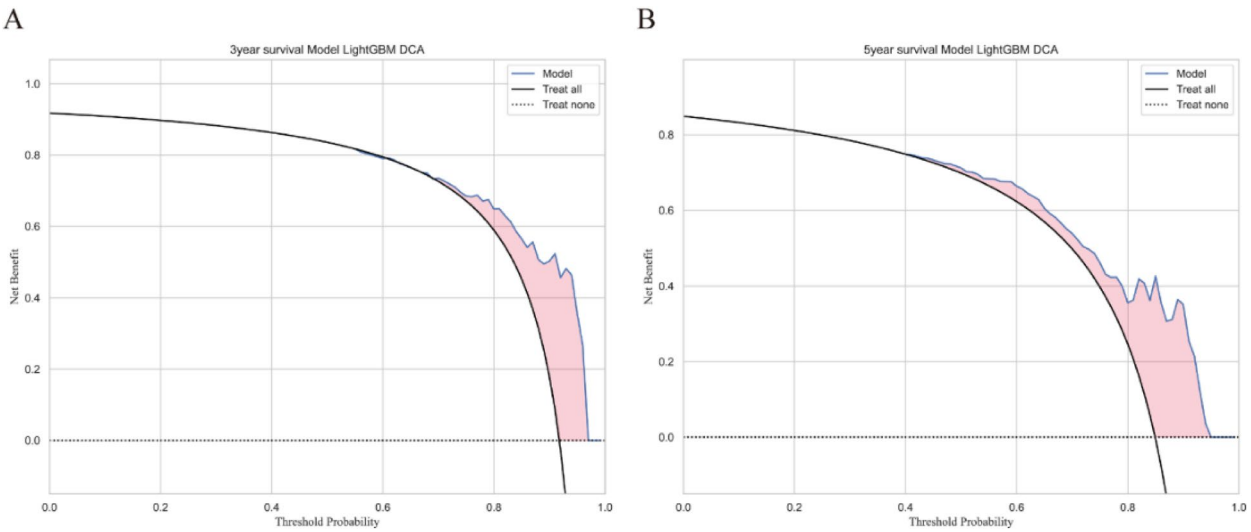

Fig. 6 Decision curve analysis of the test set of the LigntGBM model. **A** displays the DCA curve for the 3-year survival prediction model, while **B** presents the DCA curve for the 5-year survival prediction model

cancer survivors. Increasingly, evidence suggests that sarcopenia is associated with heightened mortality risks in cancer patients, although such studies are often limited to specific tumor types and demographics. For instance, a retrospective cohort study involving 3,241 non-metastatic breast cancer patients showed that, after total adjustment, sarcopenia was associated with a 41% increased risk of death (HR = 1.41; 95% CI = 1.18–1.69) [15]. Similarly, a retrospective cohort study of bladder cancer patients observed significant increases in cancer-specific mortality (HR = 2.14; $P = 0.007$) and all-cause mortality (HR = 1.93; $P = 0.004$) associated with sarcopenia [25]. Additionally, a meta-analysis on bladder cancer supported the same associations [26]. Similar outcomes were also noted in patients with non-solid tumors [27, 28]. These findings align with our results, further

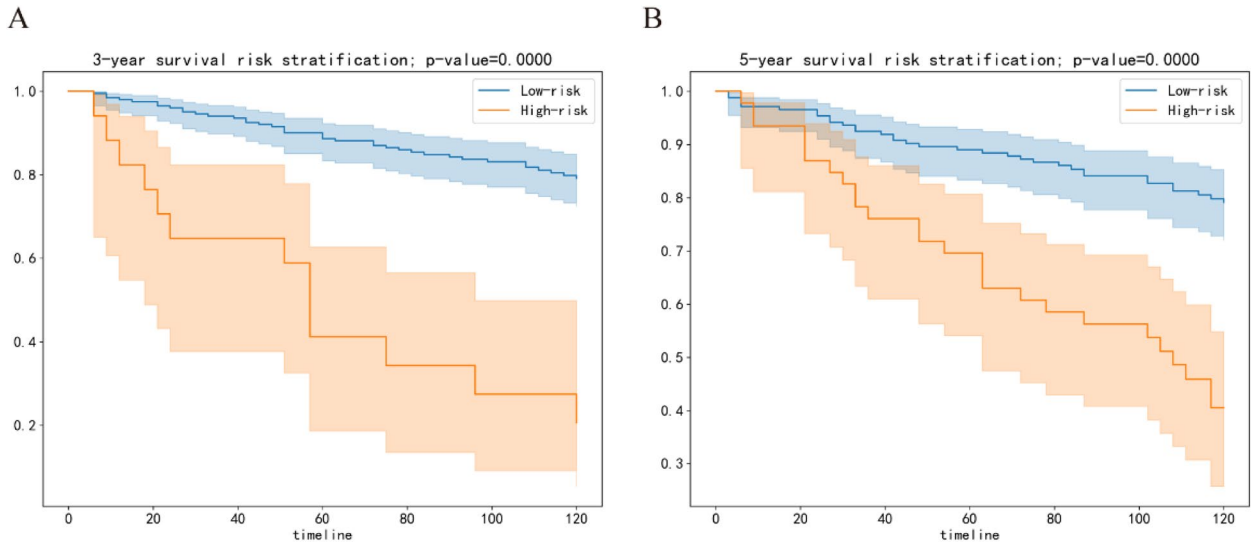

Fig. 7 Kaplan–Meier Curve Analysis Between High-risk and Low-risk Groups. Figure 7**A** and **B** illustrate the survival differences between the high-risk and low-risk groups for the 3-year and 5-year survival prediction models

emphasizing the pervasive impact of sarcopenia on survival outcomes across cancer patients, regardless of the primary tumor site or stage. Therefore, the assessment and management of sarcopenia should become a routine part of oncological care, irrespective of the cancer type or stage. This offers a new direction for optimizing cancer treatment strategies, potentially improving patient survival rates through interventions to enhance muscle condition while presenting challenges regarding healthcare resources and economic burdens. Thus, we use machine learning to construct personalized risk models, identifying cancer patients at high mortality risk over 3-year and 5-year periods. This guide targeted prevention and interventions to address these challenges.

LightGBM, an ensemble learning technique based on decision tree algorithms, has demonstrated advantages in predicting cancer patient survival rates in previous studies [29, 30]. In our performance comparison, LightGBM excelled in predicting 3-year survival rates and maintained robust performance in predicting 5-year survival rates, showcasing its strong generalizability. Hence, it was selected as the optimal model.

The exceptional performance of this model underscores the value of risk stratification, a crucial component of precision medicine, which assists physicians in more accurately predicting patient prognoses. By monitoring and intervening early on modifiable risk factors among high-risk patients, their survival rates and quality of life can be enhanced, and unnecessary medical interventions for low-risk patients can also be reduced. This strategy alleviates the public health burden and lowers overall healthcare costs, particularly in preventable medical incidents [31]. Our model effectively identifies high-risk individuals, leading us to pinpoint 10 modifiable risk factors out of the 18 identified, including sarcopenia, BMI, smoking and drinking habits, and six laboratory test indicators (uric acid, blood urea nitrogen, creatinine, ALT, Hemoglobin, and albumin). Aside from sarcopenia, the other risk factors are commonly encountered and easily obtainable in clinical settings, providing clear targets and a basis for early prevention and intervention in high-risk populations. However, sarcopenia can still be noticed. Research has shown that the increased metabolic demands of cancer and cancer-related treatments can exacerbate muscle loss [32]. Our study indicates that sarcopenia significantly increases the risk of mortality in cancer patients. Therefore, early identification and intervention for sarcopenia in cancer patients are essential.

In subgroup analyses, our study observed an interesting phenomenon. Cancer survivors with sarcopenia had a decreasing risk of all-cause and specific deaths with increasing BMI. This suggests that obesity provides additional survival benefits to such patients. This expected opposite of the association between high BMI and increased risk of death has been termed the 'obesity paradox' [33]. A recent meta-analysis found that additional obesity resulted in a lower risk of all-cause mortality in community-dwelling elderly patients with sarcopenia when categorized into sarcopenic obesity and sarcopenic non-obesity subgroups [34]. This result is more convincing than previous studies [35, 36]. In several studies investigating the correlation between obesity and the risk of death in cancer patients, researchers have employed various definitions and subgroups of obesity to eliminate the influence

of methodology on the results. However, it is noteworthy that most of the studies have observed a similar 'obesity paradox': high levels of obesity have a protective effect on the risk of death in cancer patients with sarcopenia, which is muscle loss in cancer patients [37–41].

Moreover, the reasons for this protective effect are uncertain. They may be linked to body composition, the early diagnosis and treatment of obese patients [42], and the more significant energy reserves provided by excess adipose tissue in obese cancer patients to counteract the risk of overconsumption [43]. Therefore, focusing on muscle status while considering interventions to maintain adipose tissue stability is essential in the clinical setting.

Our study boasts several strengths, including using a nationally representative cancer population database for a comprehensive analysis and adjusting multiple critical covariates to ensure the reliability of our results. Additionally, through sensitivity and subgroup analyses, we further validated the robustness and generalizability of our findings. We also developed and validated survival rate prediction models for three and five years, demonstrating excellent discriminative ability and high clinical utility. However, our research has some limitations. Firstly, although this study adjusted for a wide range of potential covariates, the possibility of residual confounding due to unmeasured factors cannot be completely ruled out. In addition, the use of self-reported data for certain variables may introduce recall bias, which could affect the reliability of the data. Secondly, as an observational study, causal relationships cannot be established. However, to minimize the potential for reverse causality, we conducted a sensitivity analysis by excluding patients who died within the first two years of follow-up. Lastly, since this study was conducted in a U.S. population and the predictive model has not yet been externally validated in an independent cohort, the generalizability of the findings to other regions or ethnic groups remains uncertain. Therefore, future prospective studies and multicenter validations in diverse populations are warranted to confirm the broader applicability of our results.

Conclusion

In summary, this study demonstrates that sarcopenia significantly increases the risk of mortality in cancer patients. Based on sarcopenia and other clinically relevant variables, we developed a survival prediction model to estimate 3-year and 5-year survival rates in cancer patients. This model effectively identifies high-risk individuals and provides new insights for advancing personalized survival assessment in cancer care.

Abbreviations

NHANES	National Health and Nutrition Examination Survey
LASSO	Least Absolute Shrinkage and Selection Operator
ML	Machine Learning
SVM	Support Vector Machine
LR	Logistic Regression
RF	Random Forest
XGBoost	Extreme gradient boosting
LightGBM	Light gradient boosting machine
NCHS	National Center for Health Statistics
DXA	Dual-energy X-ray absorptiometry
NDI	National Death Index
ASM	Appendicular skeletal muscle mass
BMI	Body mass index
MEC	Mobile Examination Centers
HR	Hazard ratio
SHRs	Subdistribution hazard ratios
AUC	Area under the curve
ROC	Receiver operating characteristic
PIR	Poverty income ratio

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12885-025-14303-9.

Additional file 1: DOC: Supplementary methods. Supplementary Table 1. Baseline characteristics according to all-cause mortality. Supplementary Table 2. Baseline characteristics of patients with different cancer types stratified by sarcopenia status. Supplementary Table 3. Univariate analysis of the COX regression model. Supplementary Table 4. Association of Sarcopenia and cancer-specific mortality via Fine & Gray competing risk model. Supplementary Table 5. Association between sarcopenia and mortality across different cancer subtypes. Supplementary Table 6. Relationship between sarcopenia and mortality in cancer patients (Excluding participants who died within 2 years of follow-up). Supplementary Table 7. Baseline characteristics of patients in machine learning model survival task. Supplementary Figure 1. Optimal cut-off values for survival risk scores.

Acknowledgements

Thanks to Jing Zhang (Shanghai Tongren Hospital) for his work on the NHANES database. His outstanding work and nhanesR package make exploring the NHANES database easier.

Authors' contributions

FC and XJD designed the study and conducted the data analysis. FC and DYP were responsible for model development and visualization. FC, XJD, YHS, YBW, RFY, and ZYH contributed to the interpretation of the results and the drafting of the manuscript. HTY and YL provided critical revisions to the manuscript. All authors read and approved the final manuscript. FC and XJD contributed equally to this work.

Funding

This study was supported by the Natural Science Foundation of Gansu Province (No.23JRRA1630, No. 24JRRA331), the Cuiying Scientific and Technological Innovation Program of Lanzhou University Second Hospital (No. CY2024-MS-B04), the Youth Science and Technology Foundation of Gansu Province (No.21JR11RA115), the Gansu Provincial Health Research Programme Project (No.GSWSKY-2019–32), the Traditional Chinese Medicine Research Project of Gansu Province (No. GZKZ-2024-26), and the Cui-Ying Scholars' Scientific Research Cultivation Programme Project of the Second Hospital of Lanzhou University (CYXZ2020-24).

Data availability

The datasets used and/or analyzed during the current study are available from the NHANES website (https://www.cdc.gov/nchs/nhanes/about_nhanes.htm).

Declarations

Ethics approval and consent to participate

The survey protocol was approved by the Research Ethics Review Board of the National Center for Health Statistics (https://www.cdc.gov/nchs/nhanes/

irba98.htm). The NCHS Research Ethics Review Board reviewed and approved NHANES, and all survey participants provided written informed consent. Therefore, no further ethical approval and informed consent were required.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Author details
[1]Department of General Surgery, Lanzhou University Second Hospital, Cui Ying Men No.80Gansu Province, Lanzhou 730030, People's Republic of China. [2]Department of Pharmaceutical, Lanzhou University Second Hospital, Cui Ying Men No.80, Lanzhou 730030, Gansu Province, People's Republic of China. [3]Department of Nuclear Medicine, Lanzhou University Second Hospital, Cui Ying Men No.80, Lanzhou 730030, Gansu Province, People's Republic of China. [4]Lanzhou University Second Hospital, Cui Ying Men No.80, Lanzhou 730030, Gansu Province, People's Republic of China. [5]School of Second Clinical Medical, Lanzhou University, Donggang West Road No. 199, Lanzhou 730030, Gansu Province, People's Republic of China. [6]Gansu High Throughput Screening and Creation Center for Health Products, School of Pharmacy, Lanzhou University, Donggang West Road No. 199, Lanzhou 730020, People's Republic of China.

Received: 12 July 2024 Accepted: 9 May 2025
Published online: 22 May 2025

References

1. Cruz-Jentoft AJ, Bahat G, Bauer J, Boirie Y, Bruyère O, Cederholm T, et al. Sarcopenia: Revised european consensus on definition and diagnosis. Age Ageing. 2019;48(1):16–31.
2. Chen LK, Woo J, Assantachai P, Auyeung TW, Chou MY, Iijima K, et al. Asian working group for sarcopenia: 2019 consensus update on sarcopenia diagnosis and treatment. J Am Med Dir Assoc. 2020;21(3):300-307.e2.
3. Yuan S, Larsson SC. Epidemiology of sarcopenia: Prevalence, risk factors, and consequences. Metab Clin Exp. 2023;144:155533.
4. Larsson L, Degens H, Li M, Salviati L, Lee Y il, Thompson W, et al. Sarcopenia: Aging-related loss of muscle mass and function. Physiol Rev. 2019;99(1):427–511.
5. Benz E, Pinel A, Guillet C, Capel F, Pereira B, De Antonio M, et al. Sarcopenia and sarcopenic obesity and mortality among older people. JAMA Netw Open. 2024;7(3):e243604.
6. Tuttle CSL, Thang LAN, Maier AB. Markers of inflammation and their association with muscle strength and mass: A systematic review and meta-analysis. Ageing Res Rev. 2020;1(64):101185.
7. Tyrovolas S, Panagiotakos D, Georgousopoulou E, Chrysohoou C, Tousoulis D, Haro JM, et al. Skeletal muscle mass in relation to 10 year cardiovascular disease incidence among middle aged and older adults: The ATTICA study. J Epidemiol Community Health. 2020;74(1):26–31.
8. Simonsen C, de Heer P, Bjerre ED, Suetta C, Hojman P, Pedersen BK, et al. Sarcopenia and postoperative complication risk in gastrointestinal surgical oncology: A meta-analysis. Ann Surg. 2018;268(1):58.
9. Feliciano EMC, Kroenke CH, Meyerhardt JA, Prado CM, Bradshaw PT, Kwan ML, et al. Association of systemic inflammation and sarcopenia with survival in nonmetastatic colorectal cancer: Results from the C SCANS study. JAMA Oncol. 2017;3(12):e172319.
10. Bray F, Laversanne M, Sung H, Ferlay J, Siegel RL, Soerjomataram I, et al. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. Ca Cancer J Clin. 2024;74(3):229–63.
11. Ryan AM, Sullivan ES. Impact of musculoskeletal degradation on cancer outcomes and strategies for management in clinical practice. Proc Nutr Soc. 2021;80(1):73–91.
12. Zhang FM, Wu HF, Shi HP, Yu Z, Zhuang CL. Sarcopenia and malignancies: Epidemiology, clinical classification and implications. Ageing Res Rev. 2023;91:102057.
13. Reisinger KW, van Vugt JLA, Tegels JJW, Snijders C, Hulsewé KWE, Hoofwijk AGM, et al. Functional compromise reflected by sarcopenia, frailty, and nutritional depletion predicts adverse postoperative outcome after colorectal cancer surgery. Ann Surg. 2015;261(2):345–52.
14. Ninomiya G, Fujii T, Yamada S, Yabusaki N, Suzuki K, Iwata N, et al. Clinical impact of sarcopenia on prognosis in pancreatic ductal adenocarcinoma: A retrospective cohort study. Int J Surg (Lond Engl). 2017;39:45–51.
15. Caan BJ, Cespedes Feliciano EM, Prado CM, Alexeeff S, Kroenke CH, Bradshaw P, et al. Association of muscle and adiposity measured by computed tomography with survival in patients with nonmetastatic breast cancer. JAMA Oncol. 2018;4(6):798–804.
16. Kourou K, Exarchos TP, Exarchos KP, Karamouzis MV, Fotiadis DI. Machine learning applications in cancer prognosis and prediction. Comput Struct Biotechnol J. 2015;13:8–17.
17. Swanson K, Wu E, Zhang A, Alizadeh AA, Zou J. From patterns to patients: Advances in clinical machine learning for cancer diagnosis, prognosis, and treatment. Cell. 2023;186(8):1772–91.
18. Huang ZN, Zhang HX, Sun YQ, Zhang XQ, Lin YF, Weng CM, et al. Multicohort study in gastric cancer to develop CT-based radiomic models to predict pathological response to neoadjuvant immunotherapy. J Transl Med. 2025;24(23):362.
19. Zhan Z, Chen B, Cheng H, Xu S, Huang C, Zhou S, et al. Identification of prognostic signatures in remnant gastric cancer through an interpretable risk model based on machine learning: A multicenter cohort study. BMC Cancer. 2024;30(24):547.
20. Chen TC, Clark J, Riddles MK, Mohadjer LK, Fakhouri THI. National health and nutrition examination survey, 2015–2018: Sample design and estimation procedures. Vital Health Stat 2. 2020;(184):1–35.
21. National Center for Health Statistics Division of Analysis and Epidemiology. Continuous NHANES Public-use Linked Mortality Files, 2019. Hyattsville, Maryland. Available from:https://www.cdc.gov/nchs/data-linkage/mortality-public.htm/10.15620/cdc:117142.
22. Buckinx F, Landi F, Cesari M, Fielding RA, Visser M, Engelke K, et al. Pitfalls in the measurement of muscle mass: A need for a reference standard. J Cachexia Sarcopenia Muscle. 2018;9(2):269–78.
23. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics (Oxford, England). 2012;28(1):112–8.
24. Camp RL, Dolled-Filhart M, Rimm DL. X-tile: A new bio-informatics tool for biomarker assessment and outcome-based cut-point optimization. Clin Cancer Res. 2004;10(21):7252–9.
25. Psutka SP, Carrasco A, Schmit GD, Moynagh MR, Boorjian SA, Frank I, et al. Sarcopenia in patients with bladder cancer undergoing radical cystectomy: Impact on cancer-specific and all-cause mortality. Cancer. 2014;120(18):2910–8.
26. Ibilibor C, Psutka SP, Herrera J, Rivero JR, Wang H, Farrell AM, et al. The association between sarcopenia and bladder cancer-specific mortality and all-cause mortality after radical cystectomy: A systematic review and meta-analysis. Arab J Urol. 2021;19(1):98–103.
27. Go SI, Kim HG, Kang MH, Park S, Lee GW. Prognostic model based on the geriatric nutritional risk index and sarcopenia in patients with diffuse large B-cell lymphoma. BMC Cancer. 2020;20(1):439.
28. Surov A, Wienke A. Sarcopenia predicts overall survival in patients with malignant hematological diseases: A meta-analysis. Clinical Nutrition (Edinburgh, Scotland). 2021;40(3):1155–60.
29. Karabacak M, Jagtiani P, Carrasquilla A, Germano IM, Margetis K. Prognosis individualized: Survival predictions for WHO grade II and III gliomas with a machine learning-based web application. npj Digital Med. 2023;6(1):200.
30. Osman MH, Mohamed RH, Sarhan HM, Park EJ, Baik SH, Lee KY, et al. Machine learning model for predicting postoperative survival of patients with colorectal cancer. Cancer Res Treat. 2022;54(2):517–24.
31. Johnson KB, Wei WQ, Weeraratne D, Frisse ME, Misulis K, Rhee K, et al. Precision medicine, AI, and the future of personalized health care. Clin Transl Sci. 2021;14(1):86–93.
32. Best TD, Roeland EJ, Horick NK, Van Seventer EE, El-Jawahri A, Troschel AS, et al. Muscle loss is associated with overall survival in patients with metastatic colorectal cancer independent of tumor mutational status and weight loss. Oncologist. 2021;26(6):e963–70.
33. Guglin M. The fiction of the obesity paradox. Eur Heart J. 2023;44(39):4199.

34. Eitmann S, Matrai P, Hegyi P, Balasko M, Eross B, Dorogi K, et al. Obesity paradox in older sarcopenic adults - a delay in aging: A systematic review and meta-analysis. Ageing Res Rev. 2024;1(93):102164.
35. Liu P, Hao Q, Hai S, Wang H, Cao L, Dong B. Sarcopenia as a predictor of all-cause mortality among community-dwelling older people: A systematic review and meta-analysis. Maturitas. 2017;1(103):16–22.
36. Zhang X, Xie X, Dou Q, Liu C, Zhang W, Yang Y, et al. Association of sarcopenic obesity with the risk of all-cause mortality among adults over a broad range of different settings: A updated meta-analysis. BMC Geriatr. 2019;19(1):183.
37. Ebadi M, Martin L, Ghosh S, Field CJ, Lehner R, Baracos VE, et al. Subcutaneous adiposity is an independent predictor of mortality in cancer patients. Br J Cancer. 2017;117(1):148–55.
38. Şahin MEH, Akbaş F, Yardimci AH, Şahin E. The effect of sarcopenia and sarcopenic obesity on survival in gastric cancer. BMC Cancer. 2023;23(1):911.
39. Lee JH, Yoon YC, Kim HS, Cha MJ, Kim JH, Kim K, et al. Obesity is associated with improved postoperative overall survival, independent of skeletal muscle mass in lung adenocarcinoma. J Cachexia Sarcopenia Muscle. 2022;13(2):1076–86.
40. Caan BJ, Meyerhardt JA, Kroenke CH, Alexeeff S, Xiao J, Weltzien E, et al. Explaining the obesity paradox: The association between body composition and colorectal cancer survival (C-SCANS study). Cancer Epidemiol Biomarkers Prev. 2017;26(7):1008–15.
41. Ozkok S, Aydin CO, Sacar DE, Catikkas NM, Erdogan T, Bozkurt ME, et al. Sarcopenic obesity versus sarcopenia alone with the use of probable sarcopenia definition for sarcopenia: Associations with frailty and physical performance. Clin Nutr. 2022;41(11):2509–16.
42. Prado CM, Gonzalez MC, Heymsfield SB. Body composition phenotypes and obesity paradox. Curr Opin Clin Nutr Metab Care. 2015;18(6):535.
43. Hughes V. The big fat truth. Nature. 2013;497(7450):428–30.

Publisher's Note