scientific reports

Check for updates

OPEN

Machine learning Nomogram for Predicting endometrial lesions after tamoxifen therapy in breast Cancer patients

Cao Shaoshan[1], Chen Niannian[2] & Ma Ying[1]✉

Objective Endometrial lesions are a frequent complication following breast cancer, and current diagnostic tools have limitations. This study aims to develop a machine learning-based nomogram model for predicting the early detection of endometrial lesions in patients. The model is designed to assess risk and facilitate individualized treatment strategies for premenopausal breast cancer patients. **Method** A retrospective study was conducted on 224 patients who underwent diagnostic curettage post-tamoxifen (TAM) therapy between November 2012 and November 2023. These patients exhibited signs of endometrial abnormalities or symptoms such as colporrhagia. Clinical data were collected and analyzed using R software (version 4.3.2) to identify factors influencing the occurrence of endometrial lesions and evaluate their predictive values. Three machine learning methods were employed to develop a risk prediction model, and their performances were compared. The best-performing model was selected to construct a nomogram of endometrial lesions. Internal validation was conducted using the bootstrap method, and the model's accuracy and fit were assessed using the concordance index (C-index) and calibration curves. **Results** Independent risk factors for endometrial lesions included ultrasound characteristics, duration of TAM therapy, presence of colporrhagia, and endometrial thickness ($P < 0.05$). Among the machine learning methods compared, the LASSO regression integrated with a multifactorial logistic regression model demonstrated strong performance, with a concordance index (C-index) of 0.874 and effective calibration (mean absolute error of conformity: 0.014). This model achieved an accuracy of 0.853 and a precision of 0.917, with a training set AUC of 0.874 (95% CI: 0.794–0.831) and a test set AUC of 0.891 (95% CI: 0.777-1.000), closely aligning the predicted risk with the actual observed risk. **Conclusion** The developed prediction model is effective in evaluating endometrial lesions in premenopausal breast cancer patients. This model offers a theoretical foundation for improving clinical predictions and devising tailored treatment strategies for this patient group.

Keywords breast cancer, tamoxifen, endometrial lesions, nomogram, prediction model

Breast cancer is among the most prevalent cancers in women worldwide, with approximately 310,000 new cases annually[1], according to the World Health Organization's International Agency for Research on Cancer (IARC) in 2024. In Asia, breast cancer exhibits a bimodal distribution, with peaks in women in their late 40s to early 50s, of whom 60% are premenopausal[2]. With the continuous development of precision medicine and treatment concept, the optimization and precision of systemic therapy has become a research hotspot in the field of breast cancer in recent years[3]. In recent years, breast cancer treatment have evolved, ranging from surgery, chemoradiotherapy, endocrine therapy to some new types of breast cancer treatments, including targeted therapy, antibody-drug conjugates[4,5]and neoadjuvant immunosuppressant therapy[6,7]. These approaches have demonstrated efficacy, with drug choices individualized based on breast cancer type. With treatment, the five-year survival rate exceeds 90%, and the ten-year survival rate surpasses 85%. New endocrine agents and antibody-drug conjugates are changing the therapeutic landscape. Approximately 75% of newly diagnosed breast cancer cases are hormone receptor-positive[8]. Endocrine therapy, primarily involving Tamoxifen (TAM), a selective estrogen receptor modulator, as a cornerstone of endocrine therapy for breast cancer, is often required for five years or more following surgery

[1]Department of Obstetrics and Gynecology, Mianyang Central Hospital, University of Electronic Science and Technology of China, Mianyang 621000, Sichuan, China. [2]School of Information Engineering, Southwest University of Science and Technology, Mianyang 621000, China. ✉email: lizza2222@126.com

and chemotherapy in premenopausal breast cancer patients[9]. Due to the receptor imbalances in endometrial tissues, TAM exerts an estrogen-like effect on the endometrium, leading to endometrial proliferation and glandular hypertrophy. This often results in abnormal glandular hyperplasia or structural changes, leading to conditions such as endometrial hyperplasia, polyps, endometrial cancer, and sarcoma[10]. Adverse reactions such as peripheral neuropathy[11]and ototoxicity[7]are also common in some cancer patients after chemotherapy and immunotherapy. Some studies have highlighted the prognostic value of the Royal Marsden Hospital Score[12]and the Neutrophil-to-Eosinophil Ratio[13] in cancer patients. With the advent of new treatment options, risk assessment methods are becoming increasingly diverse, offering enhanced opportunities to identify the optimal therapeutic strategies that maximize clinical benefits while minimizing treatment-related toxicity.

Our study focuses on the adverse effects of TAM in endocrine therapy. Some study indicate that the incidence of endometrial thickening during endocrine therapy ranges from 5 to 30%, endometrial polyps from 26 to 60%, and endometrial cancer from approximately 0.8–8%[14,15]. These rates are 2 to 7 times higher than those in the general population[10]. Additionally, Asian populations face a 2.87-fold increased risk of endometrial cancer compared to European and American groups, with approximately 20% of cases occurring in premenopausal women[8]. Several risk factors for breast cancer and endometrial carcinoma (EC) have been identified, highlighting the importance of awareness in reducing disease burden. Obesity, hypertension, and diabetes mellitus in metabolic syndrome are considered risk factors for EC and breast cancer in some previous studies. Colporrhagia frequently serves as an early symptom of endometrial cancer. Studies indicate that lifestyle-related factors, such as diet and exercise, significantly impact prognosis[16]. Consequently, identifying and addressing risk factors for endometrial lesions in younger breast cancer patients is crucial for enhancing early detection, diagnosis, and intervention.

Currently, there is no screening method that can accurately predict endometrial lesions; ultrasound remains the preferred screening tool and definitive diagnosis is established through endometrial biopsy. The risk of endometrial lesions has increased with the greater use of TAM and the longer life expectancy among breast cancer survivors, the risk of endometrial lesions is rising. Repeated ultrasounds and endometrial biopsies, which are commonly used to evaluate the need for continued endocrine therapy, are invasive and pose significant risks of complications. These procedures can adversely affect patients' mental well-being and quality of life, potentially diminishing their willingness to attend clinic visits and adhere to treatment regimens[17]. This reluctance may lead to treatment discontinuation or even tumor recurrence. Considering current research trends, early prediction of endometrial lesions remains an urgent challenge to overcome.

Although many current clinical prediction models rely primarily on simple logistic regression, the advent of artificial intelligence and machine learning has significantly enhanced prediction accuracy, particularly in fields such as medical oncology and surgery. Given the limited use of machine learning for predicting endometrial lesions in premenopausal breast cancer patients, this study aims to identify risk factors associated with endometrial lesions and to develop machine learning models to forecast their likelihood. This approach aims to improve preoperative risk assessment, offering a more tailored and theoretically informed basis for endocrine therapy in these patients.

Methods

Dataset

This study retrospectively collected data from patients who were treated with tamoxifen (TAM) after breast cancer surgery and underwent diagnostic curettage at the gynecology department of Mianyang Central Hospital between 2012 and 2023. This study was approved by Biomedical ethics committee of Mianyang Centre Hospital (No. S20240332-01). and was conducted following the ethical guidelines of the Declaration of Helsinki. Written informed consent was obtained from all participants in this study. We are committed to protecting the privacy and personal information of the participants throughout the study and to ensuring that all procedures performed in the study follow applicable ethical standards. The inclusion criteria were: (1) premenopausal status at the time of breast cancer diagnosis; (2) before breast cancer, Vaginal ultrasound was normal; (3) postoperative treatment with TAM at a dosage of 20 mg/day; (4) complete clinical case data and follow-up records. Exclusion criteria included: (1) a history of other cancers; (2) advanced or recurrent breast cancer at initial diagnosis; (3) serious underlying diseases affecting patient survival.

Patients meeting these criteria were included in the study cohort. Using a nested case-control design, patients who developed endometrial lesions were identified as the case group, while those who did not develop lesions served as the control group. Collected clinical data included age, age at breast cancer diagnosis, family history of cancer, metastasis, number of deliveries and pregnancies, weight, height, body mass index (BMI), age at menarche, breastfeeding history, endometrial thickness, presence of colporrhagia, duration of TAM therapy, previous use of sex hormones, smoking history, hypertension, diabetes mellitus, hyperlipidemia, history of fibroid tumors, endometriosis, benign breast lesions, previous endometrial lesions, ultrasound features, and endometrial biopsy results.

Development and validation of prediction model

This study followed the TRIPOD guidelines[18]. The data were initially cleaned and subjected to dimensionality reduction. The data were initially cleaned and subjected to dimensionality reduction. Univariate and multivariate analyses were conducted to identify significant factors between groups and to establish the optimal cut-off value for endometrial thickness. Three machine learning methods were employed to screen for high-risk factors: Least Absolute Shrinkage and Selection Operator(LASSO) regression combined with multifactor logistic regression, decision tree, and random forest.

A total of 224 patients were randomly divided into a training set (85%) and a test set (15%). The machine learning model was trained on the training set and validated using the test set. Model performance was assessed using the receiver operating characteristic (ROC) curve and accuracy metrics. The optimal model was selected

and internally validated with the Bootstrap resampling method (1,000 iterations). The final model was presented as a nomogram graph, and its clinical utility was with evaluated using the concordance index (C-index) and decision curve analysis (DCA).

Statistical methods

Data analysis and graphing were performed using SPSS and R software (version 4.3.2). All data were statistically described: normally distributed measurements were expressed as mean ± standard deviation ($\overline{X}\pm$ s) and compared using the t-test; non-normally distributed measurements were expressed as median [P25, P75] and compared using the rank-sum test. Categorical variables were described as constituent ratios and compared using the chi-square test or Fisher's exact test. Statistical significance was set at $P < 0.05$.

Results

Patient characteristics and endometrial biopsy results

A total of 224 premenopausal patients who underwent postoperative endocrine therapy for breast cancer were included in this study. Endometrial biopsies confirmed endometrial lesions in 98 cases, while the remaining 126 cases served as controls, resulting in a prevalence rate of 43.75% (98/224). Detailed patient characteristics are presented in Table 1.

In the lesion group, 39 cases (39.8%) had a medication duration of 2 years, 12 cases (12.2%) had a medication duration of 2–5 years, and 47 cases (48.0%) had a medication duration of more than 5 years. The mean endometrial thickness in the case group was 1.20 cm, while it was 0.76 cm in the control group, as shown in Table 2.

Correlation of clinical factors on endometrial lesions

The analysis revealed no statistically significant differences between the groups in terms of age, age at diagnosis, presence of metastasis, hypertension, diabetes mellitus, history of endometriosis, benign breast lesions, age at menarche, number of pregnancies, number of deliveries, smoking, hyperlipidemia, history of breastfeeding, family history of cancer, pathological type of breast cancer, height, weight, BMI, hormone use, history of uterine fibroids, and previous benign breast lesions ($P > 0.05$). However, significant differences were found in ultrasound characteristics, duration of TAM therapy, presence of colporrhagia, previous endometrial lesions, and endometrial thickness ($P < 0.05$). The case group showed significantly higher rates of colporrhagia, increased endometrial thickness, and longer duration of TAM therapy compared to the control group.

The mean endometrial thickness was greater in the case group (1.20 ± 0.53 cm) compared to the control group (0.76 ± 0.37 cm, $P < 0.001$), and a higher proportion of colporrhagia symptoms was observed in the case group (44.9% vs. 23.8%, $P < 0.001$). The area under the ROC curves (AUC) for each factor was as follows: ultrasound characteristics (0.770), endometrial thickness (0.749), duration of TAM (0.608), colporrhagia (0.605), and history of leiomyoma (0.560). Endometrial thickness had the largest Youden index on the ROC curve, with a cut-off value of 0.438, a sensitivity of 71.4%, and a specificity of 72.4%, corresponding to an optimal ultrasonographic diagnostic threshold of 0.825 cm for abnormal endometrium. Detailed results are presented in Table 2 and illustrated in Fig. 1.

Machine learning

The correlation heatmap revealed several linear correlations between variables (Fig. 2). To minimize interference, clinical factors were included as predictors in the machine learning model.

LASSO regression with logistic regression algorithm

In the LASSO regression to select the most predictive features. A 10-fold cross-validation was performed, resulting in the selection of four variables as independent predictors, with endometrial lesions occurrence as the dependent variable. Multifactorial binary logistic regression analysis showed that previous endometrial lesions was not statistically significant ($P > 0.05$). However, ultrasound characteristics, duration of TAM therapy, presence of colporrhagia symptoms, and endometrial thickness were identified as independent risk factors for endometrial lesions ($P < 0.05$). These factors were subsequently included in the multifactorial logistic regression model. The dataset was split into a training set ($n = 190$) and a validation set ($n = 34$) in an 8.5:1.5 ratio. After 1,000 Bootstrap self-samplings for internal validation, results showed a C-index of 0.874, indicating excellent model discrimination. The model demonstrated an accuracy of 0.853, a precision of 0.917, with an AUC of 0.874

physiology	control group($n = 126$)	Lesion group($n = 98$)
normal	126 (100)	0 (0)
Endometrial polyp	0 (0)	72 (73.5)
Endometrial hyperplasia without atypia	0 (0)	15 (15.3)
Endometrial polyp with EH	0 (0)	6 (6.1)
Atypical hyperplasia	0 (0)	2 (2)
Atypical polypoid adenomyoma	0 (0)	1 (1)
Endometrioid adenocarcinoma	0 (0)	2 (2)

Table 1. Pathological characteristics of the endometrium [cases (%)].

	control group	Lesion group	*P*-value
	126	98	
age(year)	51.68 ± 5.30	50.97 ± 5.28	0.318
Age at diagnosis of breast cancer(year)	43.91 ± 4.39	43.15 ± 5.13	0.234
gravidity	2.73 ± 1.56	2.74 ± 1.59	0.945
parity	1.00[1.00,1.00]	1.00[1.00,2.00]	0.342
menophania(year)	13.00[12.25,14.00]	13.00[12.00,14.00]	0.679
weight(kg)	57.63 ± 6.76	56.85 ± 7.54	0.416
height(m)	1.57 ± 0.05	1.57 ± 0.05	0.589
Endometrial thickness(cm)	0.76 ± 0.37	1.20 ± 0.53	< 0.001
BMI(kg/m2)	23.39 ± 2.60	23.16 ± 2.78	0.533
Ultrasonic characteristics			< 0.001
Normal	92(73.0)	28(28.6)	
Uneven echo	32(25.4)	31(31.6)	
Uterine cavity occupation	2(1.6)	35(35.7)	
Endometrium heterogeneity combined with uterine cavity occupation	0(0.0)	4(4.1)	
duration of tamoxifen therapy			0.002
Within 2 years	70(55.6)	39(39.8)	
2–5 years	24(19.0)	12(12.2)	
More than 5 years	32(25.4)	47(48.0)	
metastatic			0.061
yes	41(32.5)	20(20.4)	
no	85(67.5)	78(80.6)	
Family history of cancer			0.302
yes	18(14.3)	20(20.4)	
no	108(85.7)	78(80.6)	
breastfeeding			0.455
yes	121(96.0)	91(92.9)	
no	5(4.0)	7(7.1)	
diabetes			0.463
yes	4(3.2)	6(6.1)	
no	122(96.8)	92(93.9)	
hypertension			0.496
yes	12(9.5)	6(6.1)	
no	114(90.5)	92(93.9)	
smoking			1
yes	1(0.8)	1(1.0)	
no	125(99.2)	97(99.0)	
hyperlipemia			0.256
yes	16(12.7)	7(7.1)	
no	110(87.3)	91(92.9)	
colporrhagia			0.001
yes	30(23.8)	44(44.9)	
no	96(76.2)	54(55.1)	
hormone use			0.193
yes	7(5.6)	11(11.2)	
no	119(94.4)	87(88.8)	
leiomyoma			0.092
yes	44(34.9)	46(46.9)	
no	82(65.1)	52(53.1)	
endometriosis			0.826
yes	1(0.8)	2(2.0)	
no	125(99.2)	96(98.0)	
endometrial disease			0.008
yes	3(2.4)	12(12.2)	
Continued			

	control group	Lesion group	*P*-value
no	123(97.6)	86(87.8)	
benign lesion			1
yes	47(37.3)	36(36.7)	
no	79(62.7)	62(63.3)	
BMI group			0.843
underweight	1(0.8)	2(2.0)	
Normal	89(70.6)	66(67.3)	
overweight	32(25.4)	27(27.6)	
obesity	4(3.2)	3(3.1)	

Table 2. Baseline characteristics of premenopausal breast cancer patients [cases (%)]. Note: (a) BMI refers to body mass index, calculation formula: BMI(kg/m2) = weight(kg)/(height × height(m2)), group: underweight < 18.5, normal:18.5–23.9, overweight:24–27.9, obesity:≥28; (b) Data were expressed as (s) or n (%) or M (P25, P75), $P < 0.05$ statistically significant; nonnorm refers to data being non-normally distributed.

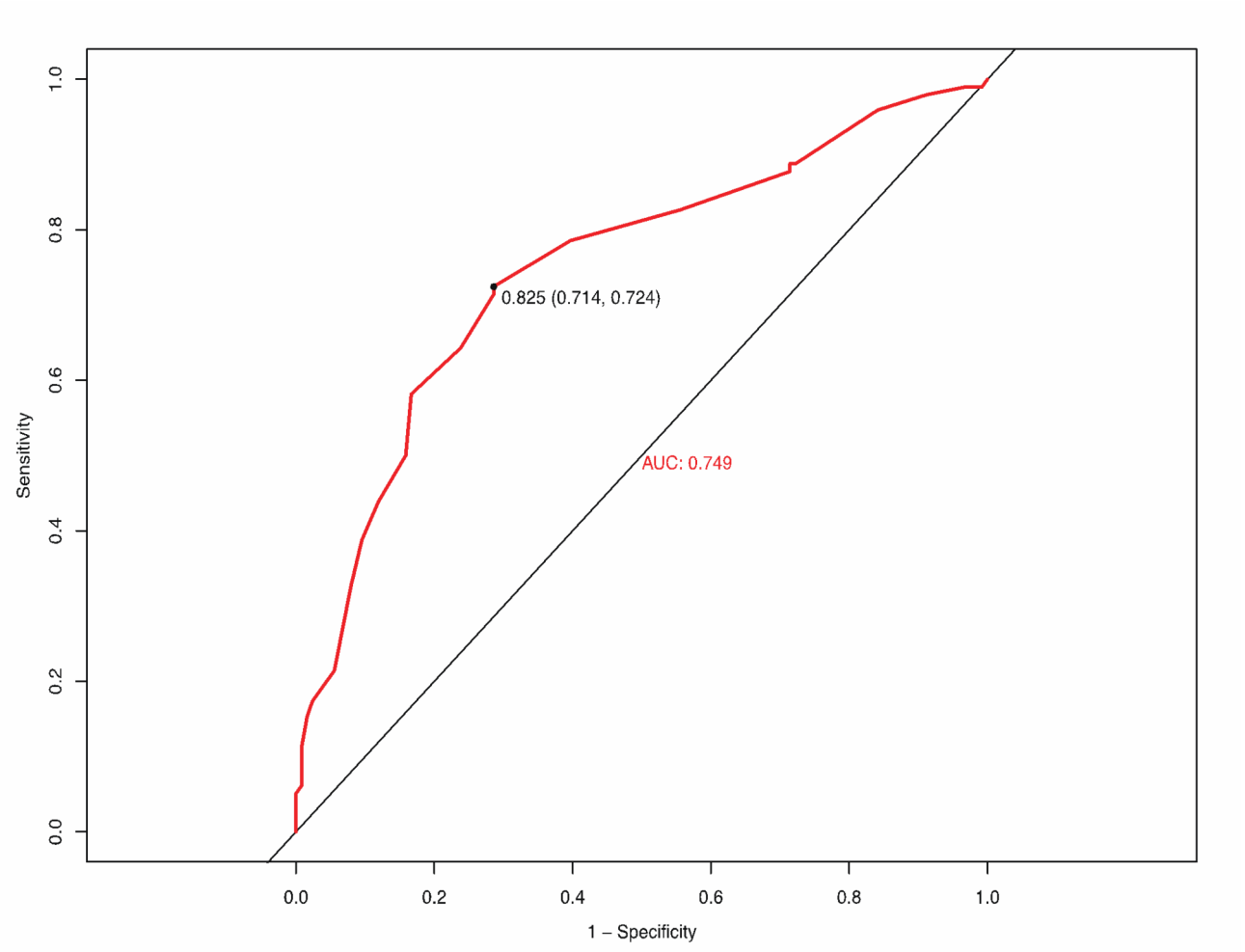

Fig. 1. ROC curve for monofactor analysis of endometrial thickness. Vertical coordinate: sensitivity, horizontal coordinate: specificity.

for training set (95% CI: 0.794–0.831) and 0.891 for a test set AUC (95% CI: 0.777–1.000). Detailed results are shown in Figs. 3 and 4a, and Table 3.

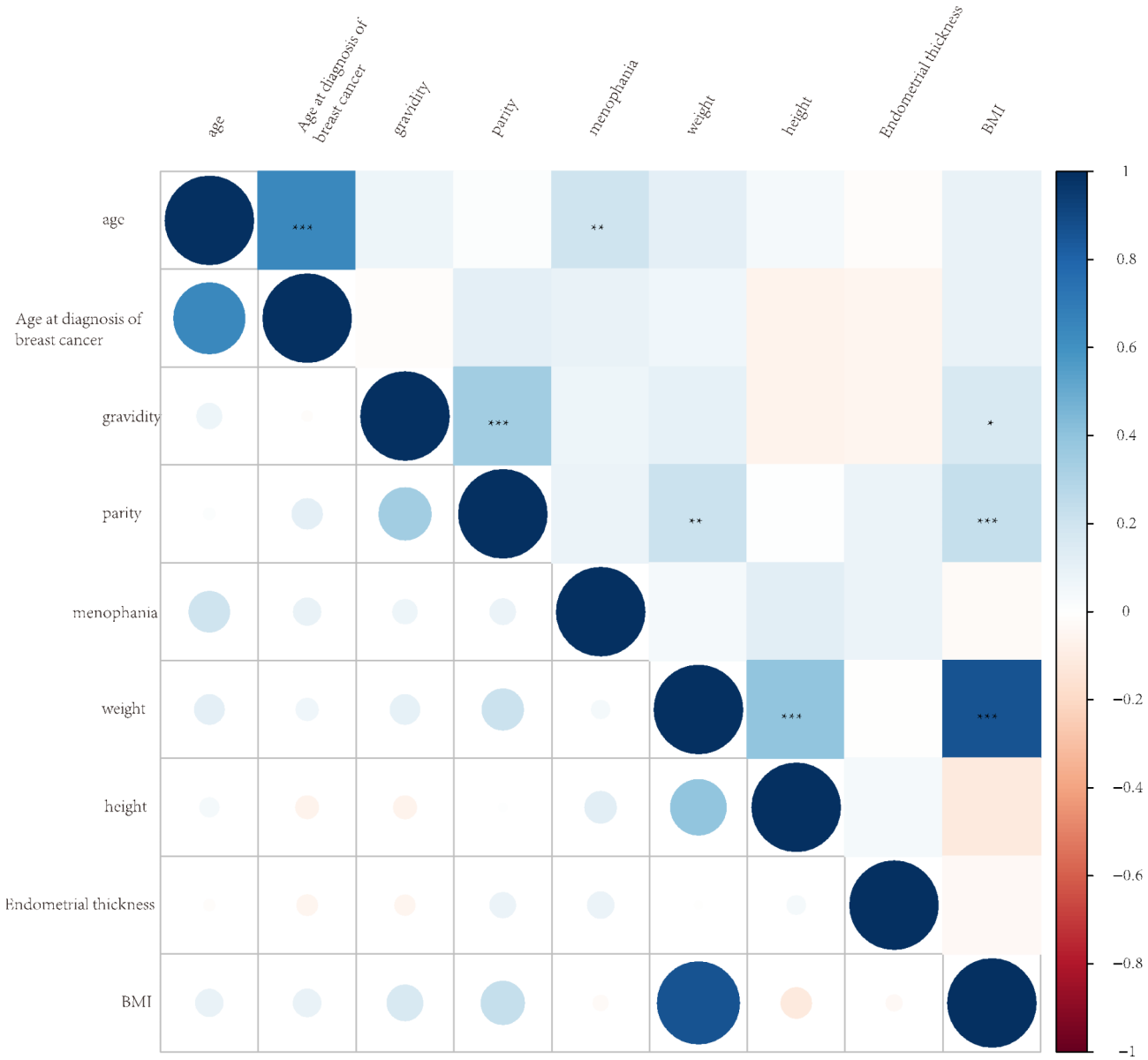

Fig. 2. Heat map of correlation between related data. * indicates correlation between data, *** means significant correlation.

Decision Tree Algorithm

A prediction model for endometrial lesions was constructed using a decision tree. The decision tree was built through recursive partitioning and pruned to avoid overfitting, with the complexity parameter(cp.) value adjusted to 0.036 to improve generalization. The average accuracy of the training set was 0.800, with an AUC of 0.888 (95% CI: 0.840–0.937). The average accuracy of the test set was 0.740 (95% CI: 0.533–0.947), with an AUC of 0.800 (95% CI: 0.640–0.960), demonstrating good prediction performance. Detailed results are illustrated in Figs. 4 and 6b.

Random Forest Algorithm

A prediction model for endometrial lesions was also constructed using a random forest approach. Five hundred trees were established, and as the number of trees increased, the out-of-bag (OOB) error rate decreased and stabilized indicating model stability. Factors were ranked in order of importance, with ultrasound characteristics, endometrial thickness, duration of TAM therapy, and colporrhagia symptoms being the most influential. The OOB error rate for the random forest training set was approximately 23.68%, with an average accuracy of 100% (95% CI: 0.981–1.000), and an AUC of 1.000 (95% CI: 1.000–1.000). The test set had an average accuracy of 0.735 (95% CI: 0.556–0.871), with an AUC of 0.784 (95% CI: 0.632–0.867), indicating good prediction performance. Detailed results are shown in Figs. 5 and 6c.

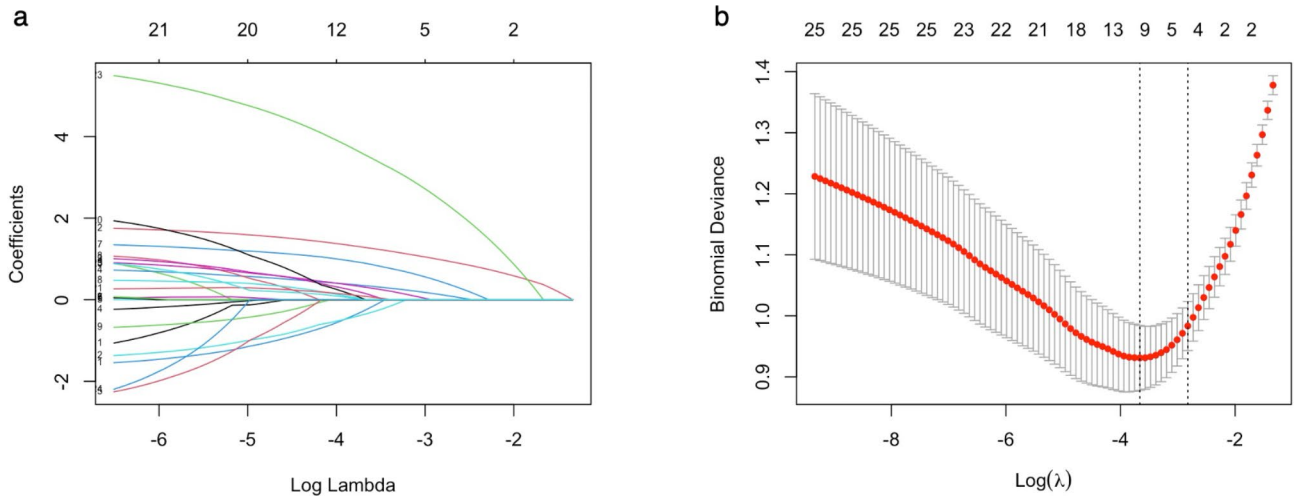

Fig. 3. Variable filtering process for Lasso regression analysis. (**a**) Cross-validation plots for selecting the optimal lambda (λ) in LASSO regressions use 10-fold cross-validation. The binomial deviation is plotted against log(λ). The left vertical dashed line indicates the value of λ associated with the minimum deviation, and the right vertical dashed line indicates the optimal value of λ determined by the minimum deviation and 1 standard deviation of the minimum deviation; (**b**) 25 characteristics were included in the LASSO regression, and a coefficient distribution plot was generated based on the log(λ) series, showing that regression coefficient estimates evolve with increasing regularization. Four non-zero coefficient variables, drug duration, endometrial thickness, and colporrhagia, were selected from the 25 variables to derive the optimal lambda.

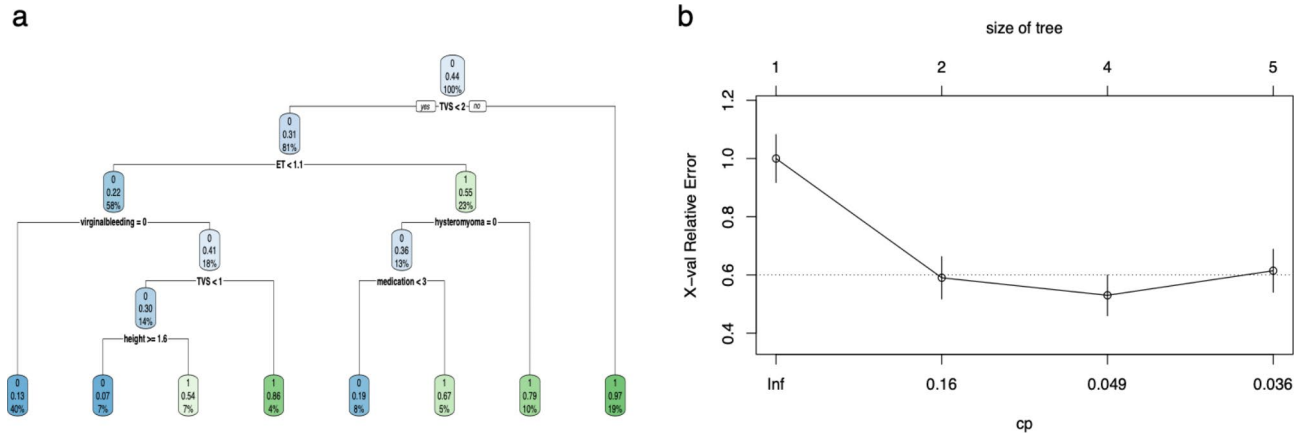

Fig. 4. Decision Tree Prediction Model after Pruning.

Model comparison and visualization

A comparison of the accuracy and AUC of the models determined that the LASSO regression combined with multifactorial logistic regression (LR) was the optimal model (see Fig. 6; Table 4).

Consequently, a nomogram graph model was constructed based on this combination for predicting the occurrence of endometrial lesions (see Fig. 7). For example, a female patient with more than five years of TAM therapy, colporrhagia, and an endometrial thickness of 1.6 cm would have corresponding scores of approximately 22.5, 17.5, and 60, respectively, totaling 100 points. This total score corresponds to an estimated probability of 80% for developing endometrial lesions. The application of the nomogram graph provides a clear and concise explanation of the model's individualized prediction for the patient.

The calibration curve suggests that the mean absolute error between predicted and actual values is 0.014, indicating that the predicted risk closely aligns with the actual risk. The Decision Curve Analysis (DCA) curve evaluates the predictive model and the concordance diagnostic test, calculating the clinical "net benefit" of the predictive model. The results show that using a column chart for predictions is valuable within a threshold probability range of 5–90% (see Figs. 8 and 9).

	Coef	S.E.	Wald Z	Df	*P*-value	OR	95%CI
Constant	−3.851	0.637	−6.04	7	< 0.001	-	-
colporrhagia	1.385	0.422	3.28	7	0.001	3.996	1.746–9.145
Endometrial thickness	1.887	0.457	4.13	7	< 0.001	3.747	2.002–7.014
duration of tamoxifen therapy				7			
2-5years: Within 2 years	0.35	0.56	0.63	-	0.532	1.419	0.474–4.253
More than 5 years: Within 2 years	0.946	0.432	2.19	-	0.029	2.575	1.104–6.006
Ultrasonic characteristics				7			
Uneven echo	1.491	0.434	3.43	-	0.001	4.44	1.897–10.394
Uterine cavity occupation	3.349	0.802	4.17	-	< 0.001	28.475	5.908–137.250
Endometrium heterogeneity combined with uterine cavity occupation	7.192	21.794	0.33	-	0.741	1328.2	3.733E-16-4.726E + 21

Table 3. Multifactorial logistic regression analysis of independent risk factors based on Lasso Regression. Note: OR: ratio of ratios; CI: confidence interval.

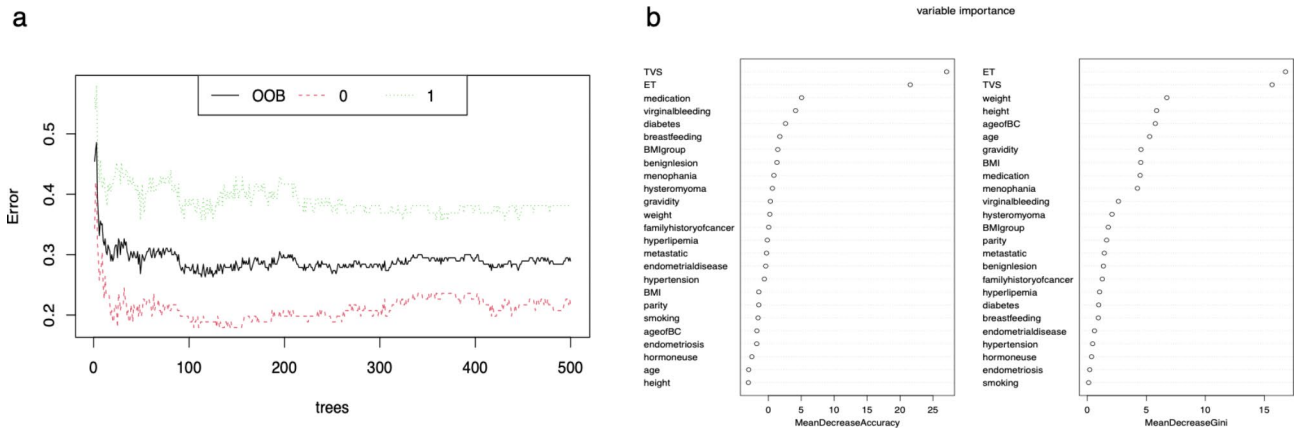

Fig. 5. Random Forest Prediction Model. (**a**) random forest model; (**b**) feature importance ranking.

Discussion

In this study, we developed and evaluated three machine learning models to accurately predict the risk of endometrial lesions in premenopausal breast cancer patients undergoing TAM therapy. The LASSO regression combined with logistic regression achieved the best predictive performance, demonstrating an accuracy of 0.853 and precision of 0.917 using four easily accessible patient features. This model exhibited high diagnostic performance with an AUC of 0.891 (95% CI: 0.777–1.000). The findings confirm that ultrasonographic features, duration of TAM therapy, endometrial thickness, and colporrhagia symptoms are significant predictors of endometrial lesions.

A national retrospective study of 102 breast cancer patients treated with TAM postoperatively found that the duration of TAM use and symptoms of abnormal colporrhagia were significant risk factors for developing endometrial lesions, consistent with our findings. Additionally, substantial epidemiologic evidence suggests that TAM is associated with an increased risk of endometrial lesions, with the risk of developing endometrial carcinoma (EC) being 1.5–6.9 times higher in a dose- and time-dependent manner[19]. The ATLAS study found that patients using TAM for 10 years had a higher cumulative risk of endometrial cancer compared to those using it for 5 years[9]. However, only 10% of patients in the ATLAS study were premenopausal, which may limit the generalizability of its findings.

Our study showed that the duration of TAM was an independent risk factor for developing endometrial lesions, consistent with previous studies[20]. Choi et al. demonstrated that benign endometrial disease incidence was highest in subjects under 40 years of age treated with TAM, significantly increasing the risk of endometrial cancer[21]. Similarly, Liu et al. found that the standardized incidence of endometrial cancer was elevated in breast cancer patients diagnosed after the age 40[22]. Younger patients treated with TAM have a higher risk of subsequent endometrial cancer, particularly those aged 40–49[23]. Bergman's study further indicated that TAM-induced endometrial cancers were more malignant and aggressive[20]. Some studies, however, have shown no correlation between TAM and endometrial lesions. For instance, Takashima[24]found no significant association between shorter TAM therapy duration and endometrial lesions. Chiofalo and Chu also reported no correlation between TAM and endometrial cancer development[23,25,26].

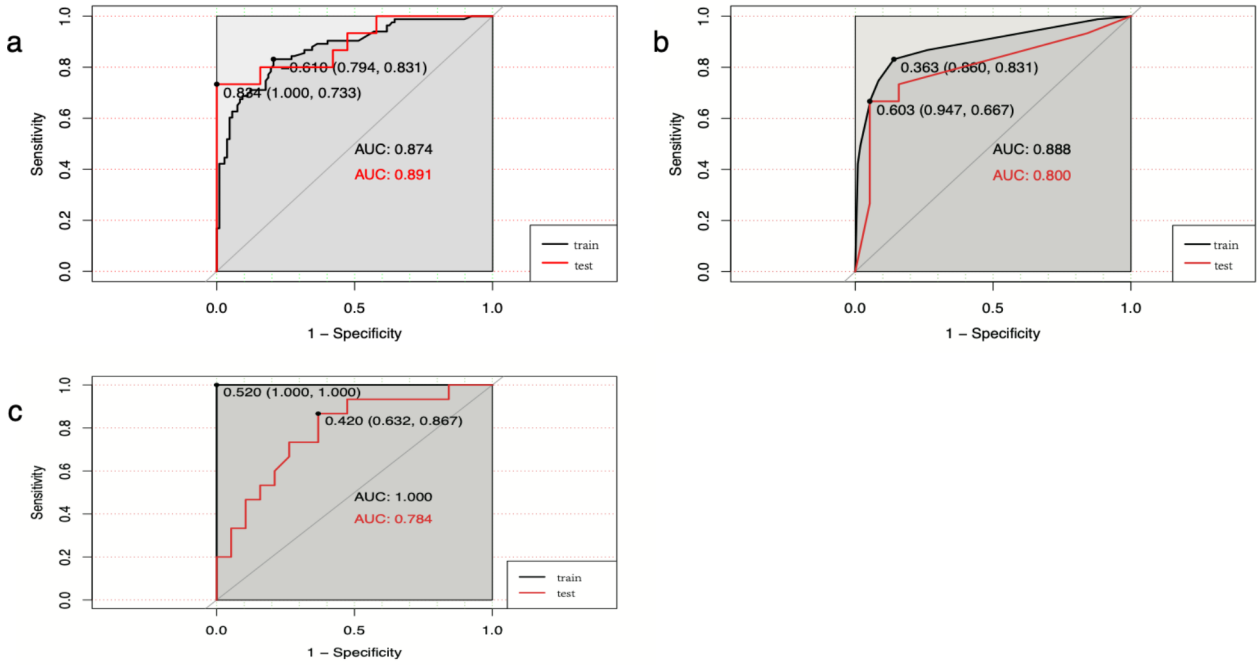

Fig. 6. Model ROC Curve. (**a**) ROC Curve of LASSO Regression with Logistic Regression Algorithm; (**b**) ROC curves for decision tree prediction model after pruning; (**c**) ROC curves for the Random Forest prediction model.

model	accuracy	AUC	sensitivity	specificity
LR	0.853	0.891	1.000	0.733
Decision tree	0.740	0.800	0.947	0.667
Random forest	0.735	0.784	0.632	0.867

Table 4. Comparison of models.

In our study, ultrasound characteristics emerged as the most important factor in predicting endometrial lesions, aligning with previous research. Ultrasound is the preferred monitoring tool, with abnormal occupancy or heterogeneous endometrial echogenicity on ultrasound increasing the likelihood of endometrial lesions and the need for endometrial biopsy. Previous NSABP studies, which included mainly postmenopausal women, suggested no additional monitoring for asymptomatic women to avoid unnecessary invasive procedures. However, this may underestimate the risk in premenopausal patients[27,28]. Young breast cancer patients undergoing prolonged TAM therapy may require closer monitoring. Endometrial screening and evaluation should be conducted before TAM treatment, followed by regular transvaginal ultrasound monitoring to enable early detection and management of endometrial lesions.

Endometrial thickness was also a significant factor in endometrial lesion occurrence, with the optimal diagnostic threshold being 0.825 cm, consistent with previous findings by Zhouqi and Burkart[2,29]. Since TAM stimulates endometrial gland hypertrophy, leading to pharmacological thickening, it is challenging to establish a TAM-related endometrial thickness threshold in young breast cancer patients.

Colporrhagia was identified as a significant risk factor. Patients with colporrhagia are more likely to develop endometrial lesions, and this symptom serves as a warning for early hospital visits, improving detection rates. However, Maria et al. found no difference in abnormal colporrhagia between the case group and patients with normal endometrium, emphasizing the need for further research[30].

Most current clinical prediction models rely on linear relationships between variables, which often limit their predictive accuracy. Machine learning applications in medicine are becoming increasingly common, providing innovative tools for clinical diagnosis and prediction. In our study, we applied machine learning techniques to visualize and predict the incidence of endometrial lesions, addressing a critical knowledge gap in evaluating premenopausal breast cancer patients undergoing endocrine therapy. By leveraging LASSO regression and multifactorial logistic regression, we mitigated the risk of overfitting and achieved validation results with an average absolute error of 0.014 between predicted and actual values. These findings highlight the potential of machine learning to revolutionize endometrial lesion prognosis, offering a significant step toward precision medicine in this field. This study holds substantial potential to improve clinical outcomes by enabling earlier

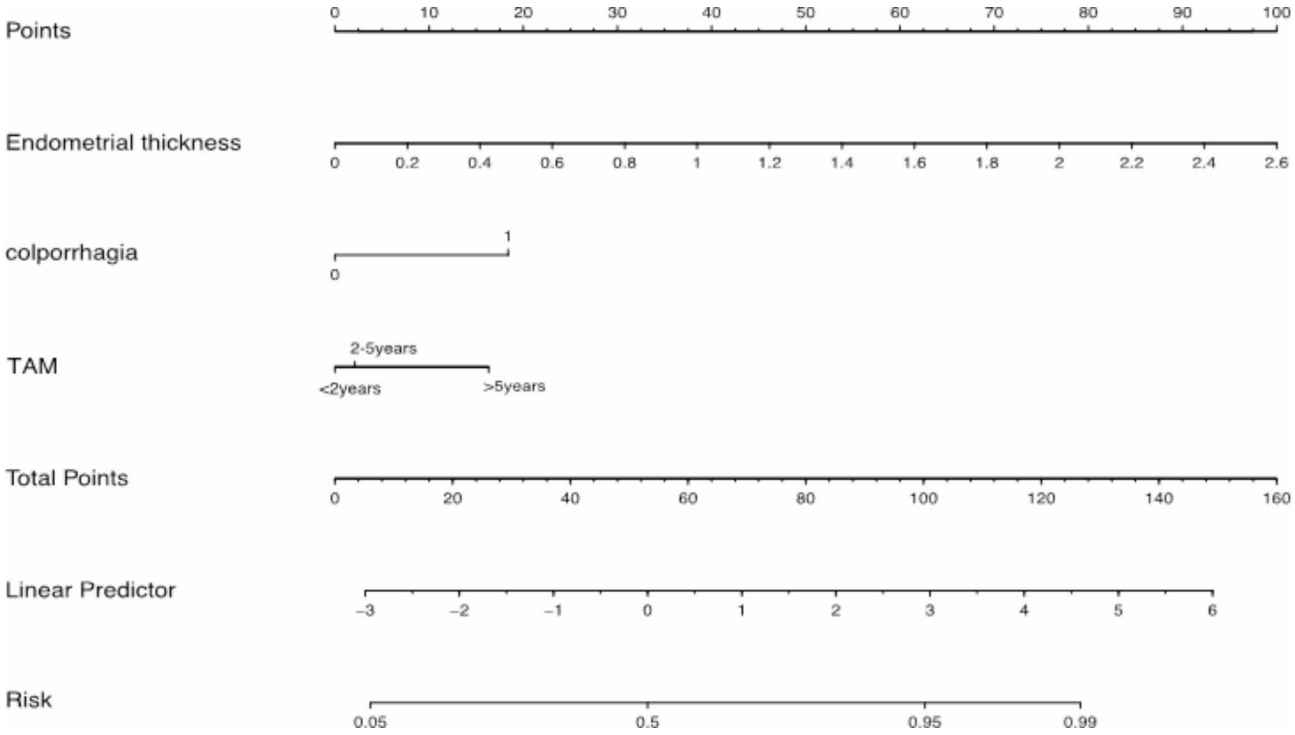

Fig. 7. Column line diagram for predicting endometrial pathology.

detection and more accurate risk prediction of endometrial lesions, particularly in vulnerable populations such as breast cancer patients undergoing hormonal treatments. By providing a theoretical foundation for developing individualized treatment strategies, this research bridges a critical gap in understanding how endocrine therapy impacts endometrial health. Researchers addressed these gaps by integrating robust statistical methods with advanced machine learning algorithms, ensuring model reliability and clinical relevance. Over the next five years, we foresee this area evolving significantly as artificial intelligence and machine learning technologies advance. Future research will likely focus on integrating multi-modal data, including imaging, genomic, and biochemical markers, to enhance the comprehensiveness of predictive models. Such developments could lead to even more accurate tools for clinical decision-making, risk stratification, and personalized treatment strategies. Furthermore, as these technologies are validated and refined, their integration into routine clinical practice will become more widespread, reducing the risk of complications and improving the overall management of breast cancer patients. In parallel, the growing emphasis on personalized medicine will likely catalyze the development of artificial intelligence tools tailored to individual patient profiles, setting new standards for treatment precision and effectiveness.

Our study also has some limitations. First, as a single-center retrospective study, our findings are inherently constrained by limited data diversity and a small sample size, which may reduce the generalizability of our results and introduce potential selection and recall biases. To mitigate these issues, future research should incorporate larger, multicenter cohorts that reflect broader population variability and improve the robustness of the findings. Second, incorporating additional objective indicators, such as hormonal profiles or advanced imaging biomarkers, could enhance the predictive accuracy of our model and establish more precise criteria for clinical use. Third, although our machine learning approach demonstrated promising results, the lack of genetic or molecular data in our analysis represents a key gap. Recent research has identified specific genes associated with breast cancer recurrence, and emerging biomarkers[31]for breast cancer prognosis could offer valuable inputs to further refine predictive models. Integrating genetic testing results and biomarker data into future studies could significantly enhance the clinical utility and precision of prediction tools[32]. Lastly, as machine learning models are only as reliable as the data they are trained on, external validation using independent datasets is essential to confirm the reproducibility of our findings. Future studies should prioritize external validation and longitudinal data to strengthen the clinical applicability of these models. Addressing these limitations will ensure that predictive models evolve into robust tools capable of supporting personalized treatment strategies and improving outcomes for breast cancer patients.

Conclusion

This study developed a predictive risk model using machine learning, with LASSO regression combined with multifactorial logistic regression demonstrating the best performance. The model identified ultrasound characteristics, TAM duration, colporrhagia, and endometrial thickness as independent risk factors for endometrial lesions in premenopausal breast cancer patients. Regular monitoring of these factors can aid in the

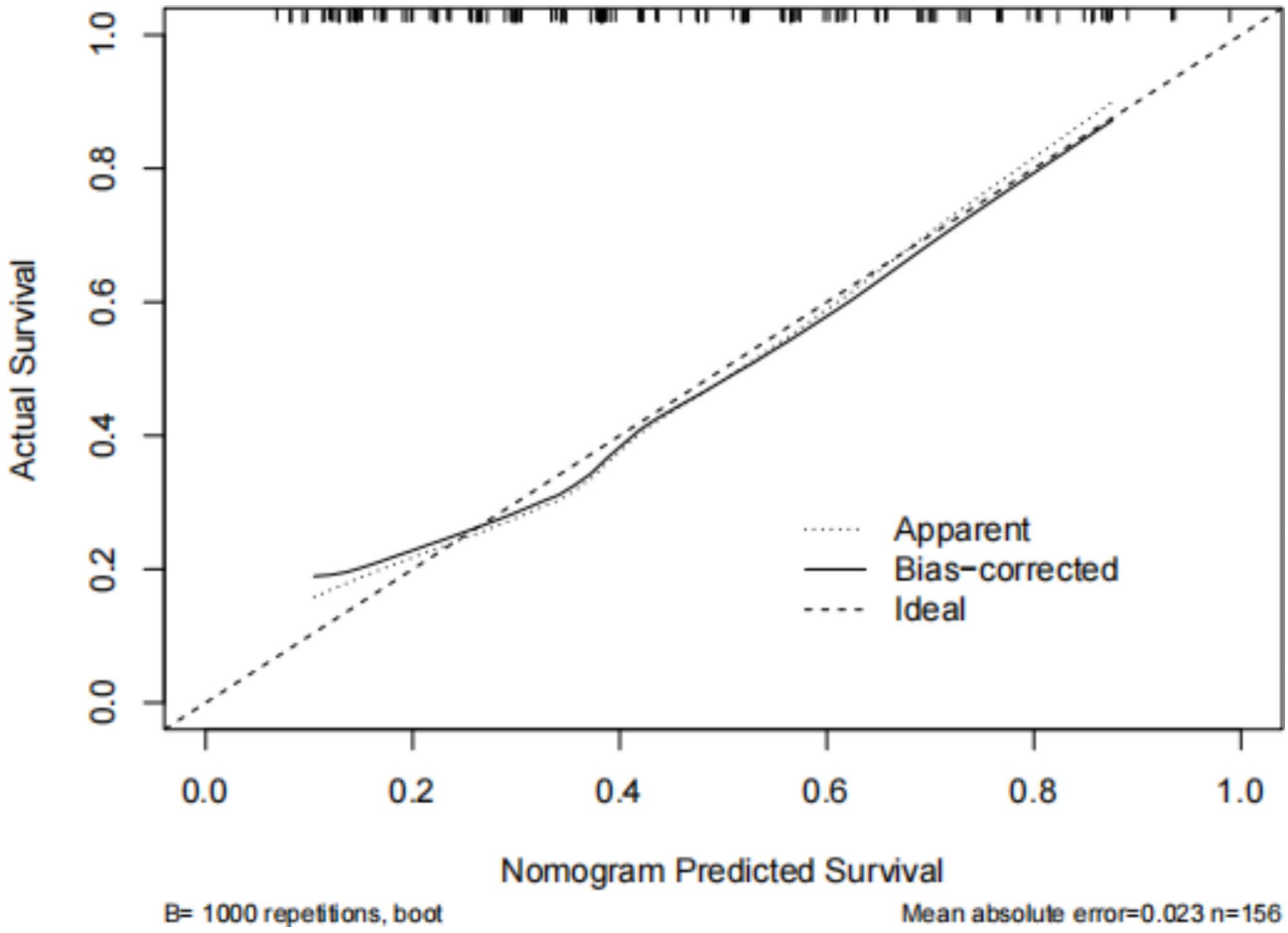

Fig. 8. Calibration curves for the column-line diagram model. Horizontal coordinate: predicted incidence of column-line plots, vertical coordinate: actual incidence. The solid black line represents the performance after internal validation by self-sampling 1000 times, the thin black dashed line represents the performance of the column-line graph, and the thick black dashed line represents the perfect prediction of the ideal model. The prediction accuracy of the column-line diagram is better when the line is closer to the thick black dashed line.

early detection and reduction of endometrial lesions, providing a foundation for evaluating endocrine therapy, endometrial monitoring during treatment, and individualized therapeutic strategies for breast cancer patients.

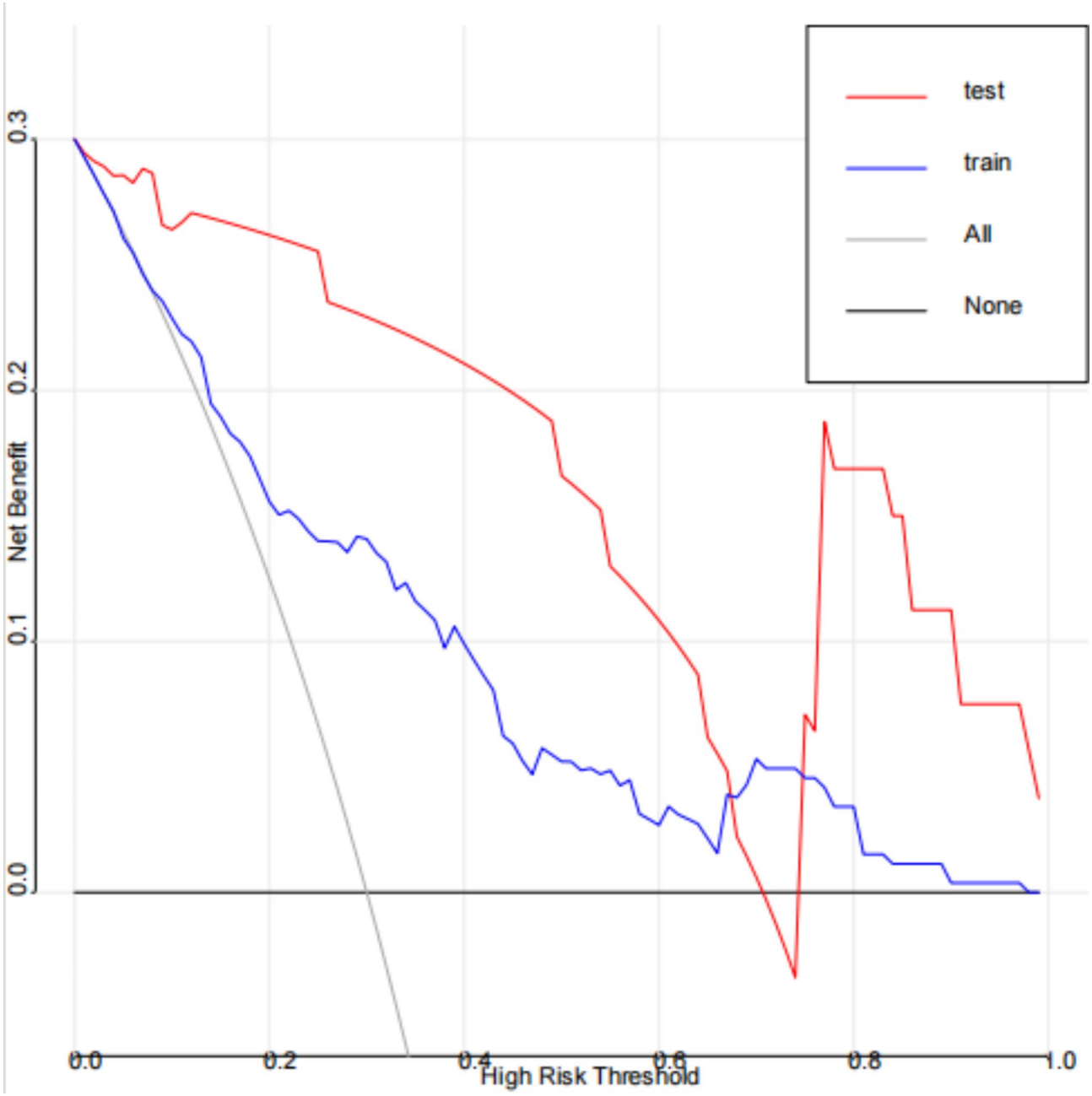

Fig. 9. Decision Curve Analysis for Column Line Charts. The horizontal coordinate is the risk threshold probability and the vertical coordinate is the net benefit after pros and cons. Blue represents the training set, red represents the test set, and gray and black represent the hypothesis that all patients have or do not develop endometriosis, respectively.

Data availability

All data are fully available without restriction. The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

Received: 10 July 2024; Accepted: 4 December 2024
Published online: 06 January 2025

References

1. SiegelRL, GiaquintoAN & Jemal, A. Cancer statistics, 2024. *CA Cancer J. Clin.* **74** (1), 12–49. https://doi.org/10.3322/caac.21820 (2024).
2. Qi, Z. H. O. U. et al. Guidelines for the management of endometriosis associated with adjuvant endocrine therapy for breast cancer (2021 edition)[J]. *Chin. J. Practical Gynaology Obstet.* **37** (8), 815–820. https://doi.org/10.19538/j.fk2021080108 (2021).

3. Huppert, L. A., Gumusay, O., Idossa, D. & Rugo, H. S. Systemic therapy for hormone receptor-positive/human epidermal growth factor receptor 2-negative early stage and metastatic breast cancer. *CA Cancer J. Clin.* **73** (5), 480–515. https://doi.org/10.3322/caac.21777 (2023).
4. Schipilliti, F. M. et al. Datopotamab deruxtecan: a novel antibody drug conjugate for triple-negative breast cancer. *Heliyon* **10** (7), e28385. https://doi.org/10.1016/j.heliyon.2024.e28385 (2024).
5. Caputo, R. et al. Sacituzumab Govitecan for the treatment of advanced triple negative breast cancer patients: a multi-center real-world analysis. *Front. Oncol.* **14**, 1362641. https://doi.org/10.3389/fonc.2024.1362641 (2024).
6. Rizzo, A. et al. KEYNOTE-522, IMpassion031 and GeparNUEVO: changing the paradigm of neoadjuvant immune checkpoint inhibitors in early triple-negative breast cancer. *Future Oncol.* **18** (18), 2301–2309. https://doi.org/10.2217/fon-2021-1647 (2022).
7. Guven, D. C. et al. Immune checkpoint inhibitor-related hearing loss: a systematic review and analysis of individual patient data. *Support Care Cancer.* **31** (11), 624. https://doi.org/10.1007/s00520-023-08083-w (2023).
8. Ghanavati, M. et al. Tamoxifen use and risk of endometrial cancer in breast cancer patients: a systematic review and dose-response meta-analysis. *Cancer Rep. (Hoboken).* **6** (4), e1806. https://doi.org/10.1002/cnr2.1806 (2023).
9. Christina Davies, H. & Pan, R. Peto. 10 vs 5 years of adjuvant tamoxifen: exclusion of 1/402 centres in ATLAS. *The Lancet*. ;389(10082):1884. doi: (2017). https://doi.org/10.1016/S0140-6736(17)31003-6
10. Emons, G., Mustea, A. & Tempfer, C. Tamoxifen and endometrial Cancer: a Janus-Headed Drug. *Cancers (Basel).* **12** (9), 2535. https://doi.org/10.3390/cancers12092535 (2020).
11. Rizzo, A. et al. Peripheral neuropathy and headache in cancer patients treated with immunotherapy and immuno-oncology combinations: the MOUSEION-02 study. *Expert Opin. Drug Metab. Toxicol.* **17** (12), 1455–1466. https://doi.org/10.1080/17425255.2021.2029405 (2021).
12. Sahin, T. K., Rizzo, A., Aksoy, S. & Guven, D. C. Prognostic significance of the Royal Marsden Hospital (RMH) score in patients with Cancer: a systematic review and Meta-analysis. *Cancers (Basel).* **16** (10), 1835. https://doi.org/10.3390/cancers16101835 (2024).
13. Sahin, T. K., Ayasun, R., Rizzo, A. & Guven, D. C. Prognostic Value of Neutrophil-to-Eosinophil ratio (NER) in Cancer: a systematic review and Meta-analysis. *Cancers (Basel).* **16** (21), 3689. https://doi.org/10.3390/cancers16213689 (2024).
14. Wenyan, T. I. A. N., Huiying, Z. H. A. N. G. & Fengxia, X. U. E. Interpretation of the Chinese expert consensus on the diagnosis and treatment of endometrial polyps (2022 edition)[J]. *J. Obstet. Gynaology.* **39** (1), 29–33 (2023).
15. Yan, G. et al. Survival nomogram for endometrial cancer with lung metastasis: a SEER database analysis. *Front. Oncol.* **12**, 978140. https://doi.org/10.3389/fonc.2022.978140 (2022).
16. Rizzo, A., Cusmai, A., Acquafredda, S., Rinaldi, L. & Palmiotti, G. Ladiratuzumab vedotin for metastatic triple negative cancer: preliminary results, key challenges, and clinical potential. *Expert Opin. Investig. Drugs.* **31** (6), 495–498. https://doi.org/10.1080/13543784.2022.2042252 (2022).
17. Park, C., Heo, J. H., Mehta, S., Han, S. & Spencer, J. C. Adherence to adjuvant endocrine therapy and survival among older women with early-stage hormone receptor-positive breast Cancer. *Clin. Drug Investig.* **43** (3), 167–176. https://doi.org/10.1007/s40261-023-01247-w (2023).
18. Collins, G. S., Reitsma, J. B., Altman, D. G. & Moons, K. G. M. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD statement. *BMJ* **350**, g7594. https://doi.org/10.1136/bmj.g7594 (2015).
19. Wijayabahu, A. T., Egan, K. M. & Yaghjyan, L. Uterine cancer in breast cancer survivors: a systematic review. *Breast Cancer Res. Treat.* **180** (1), 1–19. https://doi.org/10.1007/s10549-019-05516-1 (2020).
20. Bergman, L. et al. Risk and prognosis of endometrial cancer after tamoxifen for breast cancer. Comprehensive Cancer centres' ALERT Group. Assessment of Liver and endometrial cancer risk following tamoxifen. *Lancet* **356** (9233), 881–887. https://doi.org/10.1016/s0140-6736(00)02677-5 (2000).
21. Choi, S. et al. Risk of Endometrial Cancer and frequencies of Invasive endometrial procedures in young breast Cancer survivors treated with tamoxifen: a Nationwide Study. *Front. Oncol.* **11**, 636378. https://doi.org/10.3389/fonc.2021.636378 (2021).
22. Liu, J. et al. Elevated risks of subsequent endometrial cancer development among breast cancer survivors with different hormone receptor status: a SEER analysis. *Breast Cancer Res. Treat.* **150** (2), 439–445. https://doi.org/10.1007/s10549-015-3315-5 (2015).
23. Chu, S. C., Hsieh, C. J., Wang, T. F., Hong, M. K. & Chu, T. Y. Younger tamoxifen-treated breast cancer patients also had higher risk of endometrial cancer and the risk could be reduced by sequenced aromatase inhibitor use: a population-based study in Taiwan. *Ci Ji Yi Xue Za Zhi.* **32** (2), 175–180. https://doi.org/10.4103/tcmj.tcmj_17_19 (2020).
24. Matsuyama, Y. et al. Second cancers after adjuvant tamoxifen therapy for breast cancer in Japan. *Ann. Oncol.* **11** (12), 1537–1543. https://doi.org/10.1093/oxfordjournals.annonc.a010406 (2000).
25. AlZaabi, A. et al. Endometrial surveillance in Tamoxifen and letrozole treated breast Cancer patients. *Cureus* **13** (11), e20030. https://doi.org/10.7759/cureus.20030 (2021).
26. Chiofalo, B. et al. Hysteroscopic evaluation of endometrial changes in breast Cancer Women with or without hormone therapies: results from a large Multicenter Cohort Study. *J. Minim. Invasive Gynecol.* **27** (4), 832–839. https://doi.org/10.1016/j.jmig.2019.08.007 (2020).
27. Fisher, B. et al. Tamoxifen for prevention of breast cancer: report of the National Surgical adjuvant breast and Bowel Project P-1 study. *J. Natl. Cancer Inst.* **90** (18), 1371–1388. https://doi.org/10.1093/jnci/90.18.1371 (1998).
28. Fisher, B. et al. Endometrial cancer in tamoxifen-treated breast cancer patients: findings from the National Surgical adjuvant breast and Bowel Project (NSABP) B-14. *J. Natl. Cancer Inst.* **86** (7), 527–537. https://doi.org/10.1093/jnci/86.7.527 (1994).
29. Burkart, C. et al. [Ultrasound endometrium follow-up during tamoxifen treatment: really not reliable or useful after all?]. *Ultraschall Med.* **22** (3), 136–142. https://doi.org/10.1055/s-2001-15243 (2001).
30. Jeon, J., Kim, S. E., Lee, D. Y. & Choi, D. Factors associated with endometrial pathology during tamoxifen therapy in women with breast cancer: a retrospective analysis of 821 biopsies. *Breast Cancer Res. Treat.* **179** (1), 125–130. https://doi.org/10.1007/s10549-019-05448-w (2020).
31. Vitale, E., Rizzo, A., Santa, K. & Jirillo, E. Associations between Cancer Risk, inflammation and metabolic syndrome: a scoping review. *Biology (Basel).* **13** (5), 352. https://doi.org/10.3390/biology13050352 (2024).
32. Viscardi, G. et al. Comparative assessment of early mortality risk upon immune checkpoint inhibitors alone or in combination with other agents across solid malignancies: a systematic review and meta-analysis. *Eur. J. Cancer.* **177**, 175–185. https://doi.org/10.1016/j.ejca.2022.09.031 (2022).

Author contributions

All authors were involved in writing, all authors were involved in developing the methodology employed in the project, reviewing and editing the final draft.

Funding information

No.

Declarations

Ethics approval and consent to participate

This study was approved by Biomedical ethics committee of Mianyang Centre Hospital (No. S20240332-01). and was conducted following the ethical guidelines of the Declaration of Helsinki. Written informed consent was obtained from all participants in this study. We are committed to protecting the privacy and personal information of the participants throughout the study and to ensuring that all procedures performed in the study follow applicable ethical standards.

Competing interests

The authors declare no competing interests.
All experiments were performed in accordance with relevant named guidelines and regulations. Informed consent was obtained from the participants and/or their legal guardians for our study. Only numerical values were extracted in our study, and there is currently no publication of information about participant pictures.

Conflict of interest

None declared.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-82373-z.

Correspondence and requests for materials should be addressed to M.Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.