International Journal of Women's Health Dovepress

Open Access Full Text Article ORIGINAL RESEARCH

A Random Survival Forest Model for Predicting Residual and Recurrent High-Grade Cervical Intraepithelial Neoplasia in Premenopausal Women

Furui Zhai, Shanshan Mu, Yinghui Song, Min Zhang, Cui Zhang, Ze Lv

Gynecological Clinic, Cangzhou Central Hospital, Cangzhou, Hebei, People's Republic of China

Correspondence: Furui Zhai, Gynecological Clinic, Cangzhou Central Hospital, 16 Xinhua West Road, Cangzhou City, Hebei Province, People's Republic of China, Tel +86-0317-2075783, Email zfr860708@126.com

Purpose: Loop electrosurgical excision procedure (LEEP) for high-grade cervical intraepithelial neoplasia (CIN) carries significant risks of recurrence and persistence. This study compares the efficacy of a random survival forest (RSF) model with that of a conventional Cox regression model for predicting residual and recurrent high-grade CIN in premenopausal women after LEEP.
Methods: Data from 458 premenopausal women treated for CIN2/3 at our hospital between 2016 and 2020 were analyzed. The RSF model incorporated demographic, pathological, and treatment-related variables. Feature selection utilizing LASSO and three other algorithms was performed to enhance the RSF model, which was further compared to a Cox regression model. Model performance was assessed using area under the curve (AUC), out-of-bag (OOB) error rates, and SHAP values to interpret predictor importance.
Results: The RSF model showed superior performance compared to the Cox regression model, with AUC values of 0.767–0.901 and peak predictive performance at 36 months post-LEEP. In contrast, the highest AUC achieved by Cox regression was 0.880. The RSF model also exhibited relatively lower OOB error rates, indicating better generalizability. Moreover, SHAP value analysis identified margin status and CIN severity as the most prominent predictors that directly affected risk predictions. Lastly, an online tool providing real-time predictions in clinical settings was successfully implemented using the RSF model.
Conclusion: The RSF model outperformed the traditional Cox regression model in predicting residual and recurrent high-grade CIN risks post-LEEP. This model may be a more accurate clinical tool that facilitates improved personalized care and early interventions in gynecological oncology.

Keywords: cervical intraepithelial neoplasia, residual/recurrent, random survival forest, Cox regression, premenopausal women

Introduction

Cervical cancer is the fourth most prevalent cancer in women worldwide and a leading cause of female cancer-related deaths.[1] High-grade cervical intraepithelial neoplasia (CIN) is a precancerous condition that requires timely treatment to prevent progression to cancer,[2] with loop electrosurgical excision procedure (LEEP) being the preferred intervention method.[3] Despite undergoing successful initial treatment, many patients are at risk of experiencing persistent or recurrent lesions, and the identification of such high-risk individuals can pose a remarkable clinical challenge. Therefore, accurately predicting residual and recurrent high-grade CIN is essential for optimizing post-treatment strategies and improving patient outcomes, especially in premenopausal women requiring fertility preservation.

The incidence of high-grade CIN and cervical cancer from 1990 to 2019 has notably increased among women of reproductive age.[4] This rising trend is closely linked to heightened estrogen levels in premenopausal women, which can significantly increase the risk of HPV infection—a key contributor to cervical disease progression.[5] Although the direct comparisons of residual disease and recurrence rates after conization between premenopausal and post-menopausal women are limited, existing evidence suggests that premenopausal women may have higher residual and recurrence rates.[3,6,7] This

Received: 13 August 2024
Accepted: 24 October 2024
Published: 30 October 2024

finding underscores the urgent need for targeted research to formulate personalized treatment approaches that balance cancer prevention with fertility preservation in this vulnerable population.[8]

Various factors, such as age, HPV type, lesion size, surgical margin status, depth of invasion, and hormonal influences, can influence the likelihood of residual and recurrent disease; however, understanding the complex interactions among these factors remains difficult.[9–12] Although traditional predictive models, including Cox regression, have been employed to estimate these risks, they have certain inherent limitations.

Cox proportional hazards models are widely utilized in survival analysis and prognosis prediction, including in cervical cancer research.[13–16] However, these models rely on assumptions of linear covariate relationships and proportional hazards, which may not adequately capture the complex, non-linear interactions between clinical, pathological, and demographic factors associated with high-grade CIN recurrence.[17] In contrast, machine learning models, such as the random survival forest (RSF), offer a more flexible approach to survival analysis. The RSF model is designed to handle non-linear relationships and high-dimensional data by constructing an ensemble of decision trees. This approach captures complex interactions between variables while reducing overfitting and managing missing data or censoring. Its flexibility allows it to work with various types of predictors, making it well-suited for survival analysis.[18]

Recent studies have shown that the RSF model provides superior predictive performance compared to Cox regression, particularly in diseases where non-linear interactions have a pronounced role.[19] Therefore, the RSF model can be used to better capture the intricacies of factors influencing residual and recurrent high-grade CIN in premenopausal women. In this study, we applied the RSF model to develop a more accurate predictive tool as well as incorporated a wide range of clinical, pathological, and demographic variables to improve personalized care and post-treatment outcomes.

Methods

Study Population

This study analyzed the clinical and pathological data of premenopausal women with CIN2/3 who were treated at our hospital between January 2016 and December 2020. All patients underwent LEEP and were followed up until December 2021. The study was approved and ethically monitored by the hospital's ethics committee. Collected patient data included demographics, reproductive history, menopausal status, ThinPrep cytologic test (TCT) results, HPV classification, cervical lesion extent, glandular involvement, and initial LEEP margin status.

Eligibility Criteria

The study included premenopausal women aged 20–50 years who underwent LEEP following a diagnosis of CIN2 or worse and consented to follow-up assessment. Patients were excluded if they were post-menopausal, had concurrent gynecological or severe systemic diseases, liver or kidney impairment, prior total hysterectomy, post-operative invasive cervical cancer, previous cervical pathologies, hormone replacement therapy, or acute infections, or were pregnant.[15,20]

Critical Definitions

Specialized gynecologists performed cervical surgeries involving the excision of a cone-shaped section from the transformation zone, the primary site for CIN. The depth and edges of the excision were modified to ensure effective lesion removal while preserving cervical integrity, as assessed via colposcopy. Residual or recurrent disease was primarily detected through the histopathological analysis of the biopsies of the lesions identified within or after 1-year post-LEEP procedure. The residual and recurrent conditions were grouped for further analysis due to their comparable clinical significance.

Follow-Up Protocol

Patients were followed up semi-annually for 2 years, followed by an annual examination. In patients with positive HPV findings, additional colposcopy and biopsy examinations were required. Moreover, histological assessment was performed during these follow-ups to grade the most severe abnormalities that were identified. All procedures were conducted under the supervision of expert gynecologists and confirmed by pathologists. Follow-up was continued from the conization procedure to the study's conclusion until the detection of residual/recurrent CIN or patient dropout or death.

Feature Selection Methods

Various algorithms were utilized to determine the predictor of residual and recurrent high-grade CIN following the conization method. Feature selection was performed using advanced algorithms, including the least absolute shrinkage and selection operator (LASSO) regression to minimize overfitting by applying penalties to regression coefficients,[21] the Boruta algorithm (version 8.0.0) to identify crucial classification features,[22] support vector machine-recursive feature elimination with cross-validation (SVM-RFE-CV) to optimize model accuracy by pruning insignificant features,[23] and the ReliefF method to highlight meaningful feature interactions.[24]

Model Development and Evaluation

The RSF model was developed utilizing clinical data from premenopausal women with high-grade CIN, with the data divided into 80% training and 20% testing sets. The RSF method was implemented using randomForestSRC 3.2.3. Model performance was internally validated through out-of-bag (OOB) error rates and externally validated via receiver operating characteristic (ROC) curves. Decision curve analysis (DCA) was also conducted to assess clinical utility across various thresholds by employing rms 6.7.1 and survminer 0.4.9, along with cross-validation checks to ensure model stability and generalizability. The Cox regression model was applied to predict residual/recurrence occurrence post-conization, and its performance was evaluated at multiple intervals through pROC 1.18.5. Calibration curves were also used to verify prediction accuracy against actual outcomes at 48 months, with the concordance index from survminer 0.4.9 reflecting overall predictive performance. Lastly, multivariate Cox regression was conducted using the significant predictors, with forest plots (constructed utilizing forestploter 1.1.1) for visualizing the results.

Interpretation of Random Forest Survival Analysis

Shapley Additive Explanations (SHAP) values obtained from the SHAP library (version 0.43.0) in Python were employed to measure the influence of predictors on survival results. SHAP values, which are based on cooperative game theory, precisely illustrate the influence of each predictor and thus offer a clear understanding of their effects on the model's predictions. Consequently, SHAP plots can be used to visually represent the significance of the variables on the outcomes, improving the interpretability of the model and validating its utility in clinical scenarios.

Risk Stratification and Survival Evaluation

Patients were categorized into high- and low-risk groups according to the threshold determined by the RSF model, which was optimized to differentiate survival outcomes in the training and testing datasets. Moreover, Kaplan–Meier estimates were utilized to plot survival probabilities for each group, and the Log rank test was employed to evaluate the significant differences in survival outcomes. Finally, the model's performance and consistency were validated using an independent test set.

Statistical Methods

Statistical analyses were performed using R (version 4.2.3) and Python's Scikit-Learn library (version 1.1.3). Differences between the training and testing sets were assessed using the Wilcoxon test for continuous variables and the χ^2 or Fisher's exact tests for categorical variables. Statistical significance was set at a two-tailed p-value of <0.05.

Results

Comparison of Patient Characteristics

A total of 458 premenopausal women were included in this study, among which 383 had no residual or recurrent CIN, while 75 experienced residual disease or recurrence. Age distribution between the two groups was similar, with a mean age of 37.11 ± 6.32 years in the non-recurrent group and 36.12 ± 6.91 years in the recurrent group (P = 0.220). Pregnancy and parity were also not significantly different between the two groups (P > 0.05). However, a notable variation was observed in the TCT results, where the recurrence group showed a higher prevalence of high-grade lesions than the non-recurrence group (64% vs 35.77%; P < 0.001). Furthermore, the recurrence group had a higher incidence than the non-recurrence group in terms of high-risk HPV types 16/18 (78.66% vs 52.48%; P < 0.001) and CIN3 (69.33% vs 25.06%; P < 0.001). Positive surgical margins were also more frequently reported in the recurrence group than in the non-recurrence group (72.00% vs 20.10%; P < 0.001). However, glandular involvement did not differ significantly between the two groups (P > 0.05). Detailed comparisons of the patient characteristics are provided in Table 1.

Feature Selection Process

LASSO regression identified eight significant predictors (Figure 1A), and Boruta analysis highlighted the importance of "HPV", "degree of CIN", "TCT results", and "margin status" (Figure 1B). In the case of SVM-RFE-CV, "HPV", "degree of CIN", "glandular involvement", and "margin status" were selected as critical features, while ReliefF emphasized the relevance of "parity", "TCT results", and "pregnancy" as crucial factors (Figure 1C and D). Table 2 compares the top predictors from each method, showing consistency across models. The integration of the results of these methods led to the identification of six critical variables for the predictive model: "margin status", "degree of CIN", "glandular involvement", "parity", "TCT results", and "HPV". This rigorous feature selection process resulted in a predictive model that was highly relevant for predicting disease outcomes in premenopausal women with high-grade CIN.

Model Performance and Evaluation

This study utilized the RSF model to predict residual disease and recurrence in premenopausal women following conization for high-grade CIN. Patients were divided into training and testing sets at a 4:1 ratio. Subsequent ROC curve (Figure 2A and B) analysis demonstrated strong predictive performance, with the training and testing groups achieving area under curve (AUC) values of 0.886, 0.862, 0.901, and 0.839 and 0.858, 0.790, 0.779, and 0.767 at 12, 24, 36, and 48 months, respectively. Cumulative survival curves (Figure 2C) highlighted notable differences in survival probabilities between the low- and high-risk groups. Additionally, consistent OOB error rates (Figure 2D) confirmed

Table 1 Baseline Characteristics of the Population

Patient Characteristic	No Residual/Residual CIN (n=383)	Residual/Residual CIN (n=75)	*P-value*
Age (years)	37.11 ± 6.32	36.12 ± 6.91	0.220
Pregnancy, n (%)			0.641
<3	193 (50.39)	40 (53.33)	
≥3	190 (49.60)	35 (46.66)	
Parity, n (%)			0.068
<2	136 (35.50)	35 (46.66)	
≥2	247 (64.49)	40 (53.33)	
TCT, n (%)			<0.001
<ASC-H	246 (64.23)	27 (36.00)	
≥ASC-H	137 (35.77)	48 (64.00)	
HPV, n (%)			<0.001
No HR-HPV	9 (2.35)	2 (2.66)	
HPV16/18	201 (52.48)	59 (78.66)	
Other HR HPV	173 (45.17)	14 (18.66)	
Degrees of CIN, n (%)			<0.001
CIN2	287 (74.93)	23 (30.66)	
CIN3	96 (25.06)	52 (69.33)	
Glandular involvement, n (%)			0.238
No	252 (65.79)	44 (58.66)	
Yes	131 (34.20)	31 (41.33)	
Margin status, n (%)			<0.001
Negative	306 (79.89)	21 (28.00)	
Positive	77 (20.10)	54 (72.00)	

Notes: Data are shown as mean±standard deviation or median (interquartile range) or percentage.
Abbreviations: CIN, cervical intraepithelial neoplasia; TCT, ThinPrep cytological test; ASC-H, atypical squamous cells cannot exclude high grade squamous intraepithelial lesion; HR-HPV, high-risk human papilloma virus. P value< 0.05 was considered significant.

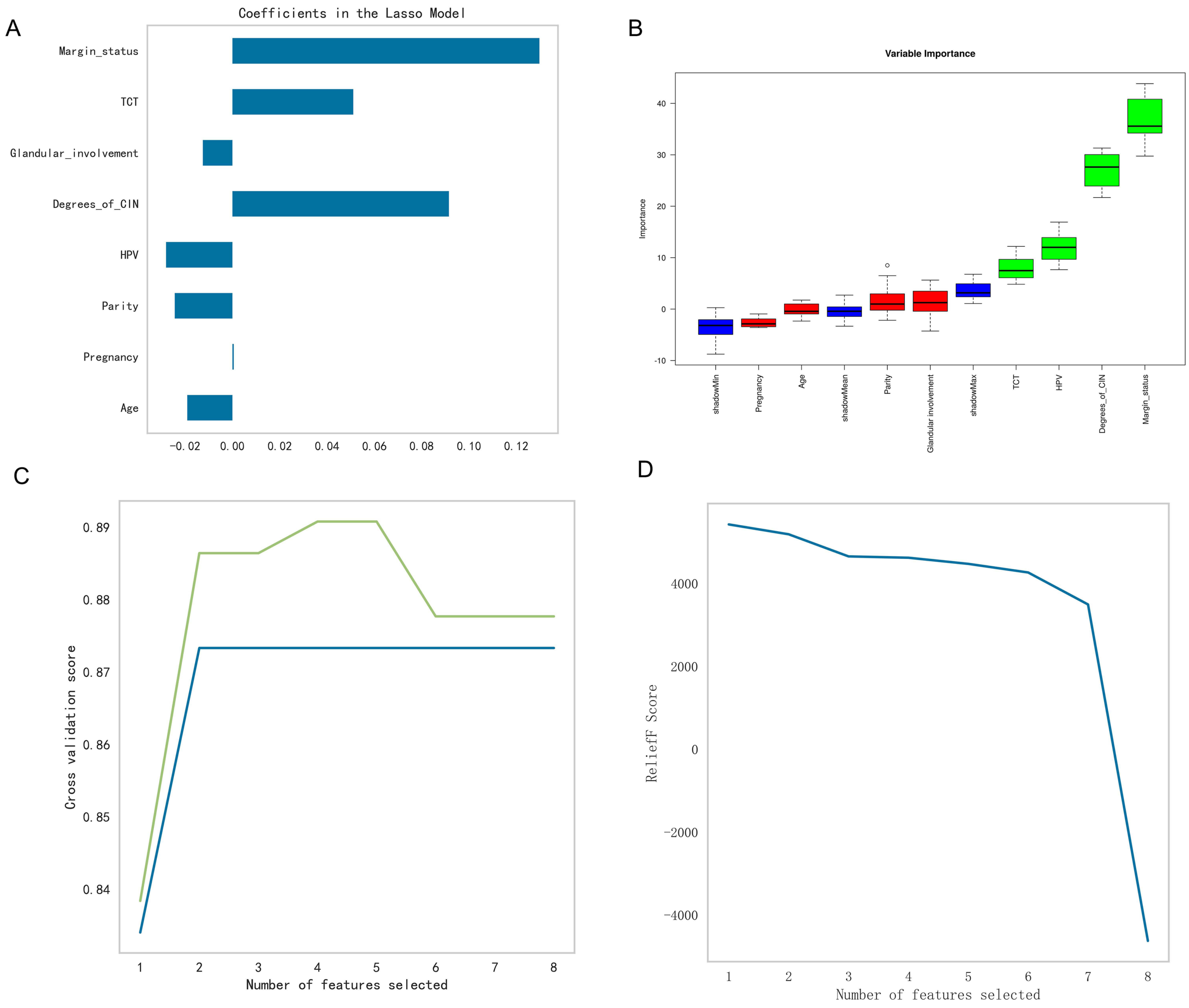


Figure 1 Feature selection analysis for the predictive modeling of high-grade CIN following conization. (**A**) A bar chart displaying the coefficients of the features selected by the LASSO regression algorithm, where the bar lengths indicate the coefficient magnitudes. (**B**) A boxplot showing the distribution of importance scores for the variables identified by the Boruta algorithm, with ranking based on their median importance score. (**C**) A line graph illustrating the model's cross-validation scores as a function of the number of selected features, along with an annotation for the optimal number of features. (**D**) A plot depicting the importance scores of the ReliefF features corresponding with the increasing number of selected features.

model stability, while the cumulative proportional score (CPS) trends demonstrated consistent performance across varied percentiles (Figure 2E). A scatter plot (Figure 2F) indicated good agreement between the predicted and actual outcomes, and DCA (Figure 2G) validated the clinical utility of the model.

Table 2 Comparison of Top Predictors Selected by Feature Selection Methods

Predictor	LASSO	Boruta	SVM-RFE-CV	ReliefF
Margin Status	√	√	√	
Degree of CIN	√	√	√	
HPV Status	√	√	√	
TCT Results	√	√		
Glandular Involvement	√		√	√
Parity	√			√
Pregnancy	√			√

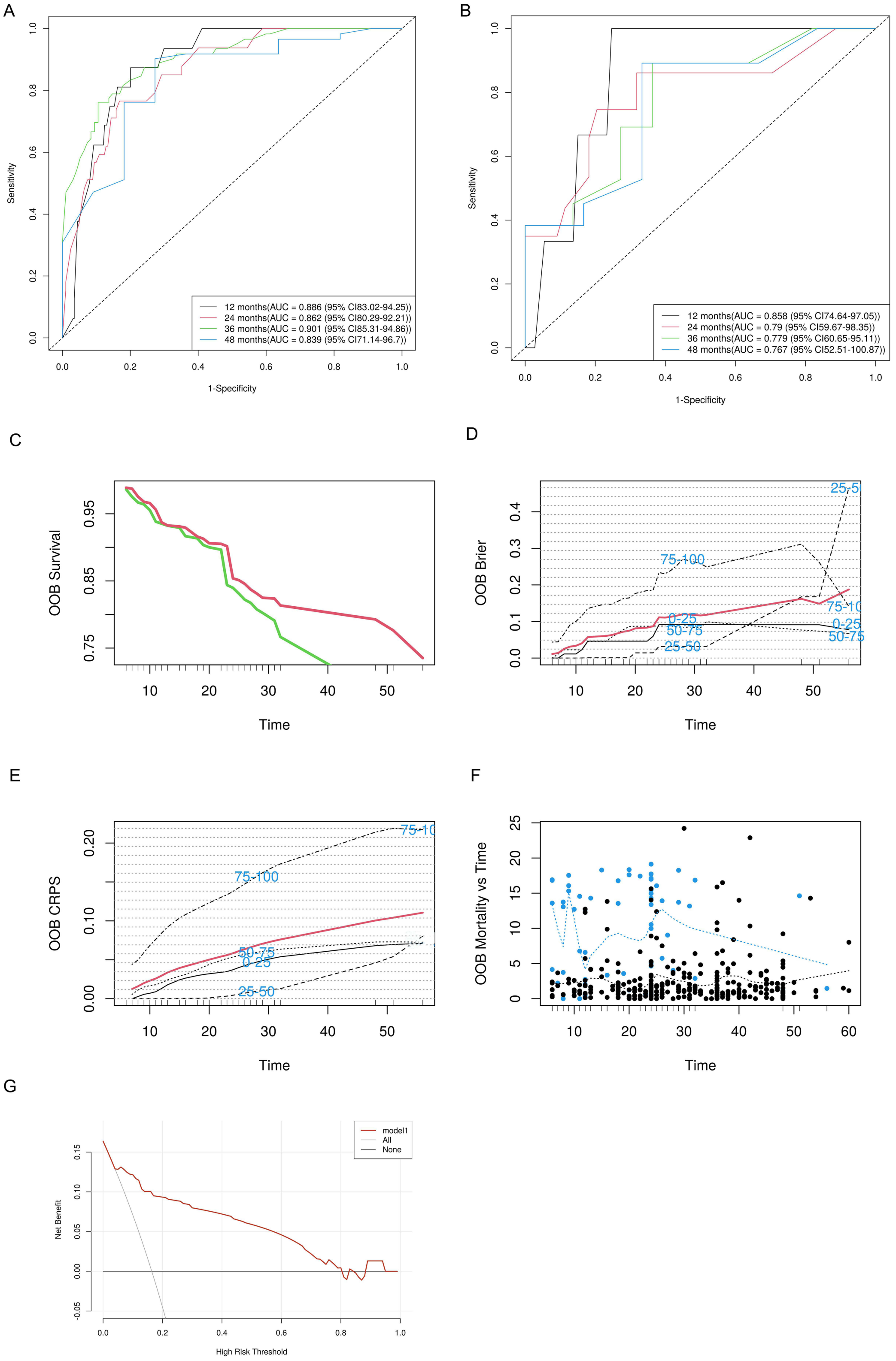


Figure 2 Evaluation of the random forest model for predicting residual disease and recurrence in premenopausal women who underwent conization for high-grade CIN. ROC curves of the random forest model at follow-up intervals for the training (**A**) and testing (**B**) groups. (**C**) Cumulative survival curves representing the survival differences between the low- and high-risk groups. (**D**) Time series of the OOB error rates showing stabilization as the model iterates. (**E**) Cumulative proportional score (CPS) providing a time series representation for various percentiles. (**F**) A scatter plot with a blue dashed trend line demonstrating good agreement between the predicted and actual outcomes. (**G**) Decision curve analysis of the random forest model in the testing set.

The Cox regression model provided additional valuable insights, yielding a consistent AUC value of 0.83 in the training set and AUC values ranging from 0.78 to 0.87 in the testing set across different follow-up times (Figure 3A and B). A slight decrease (AUC = 0.78) at 48 months suggested certain limitations in the accuracy of long-term predictions. Nevertheless, a calibration curve at 48 months (Figure 3C) and a high concordance index confirmed a strong correlation between the predicted and actual outcomes. Key predictors, such as margin status (HR = 7.04, $p < 0.01$) and degree of CIN (HR = 3.13, $p < 0.01$), were prominently demonstrated in a forest plot (Figure 3D), underscoring their significant influence on residual disease and recurrence.

Comparison of Model Performance

As depicted in Table 3, the RSF model consistently outperformed the Cox regression model in the training set. In particular, the RSF model achieved AUC values of 0.886, 0.862, 0.901, and 0.839 from 1 to 4 years, respectively. In contrast, the Cox

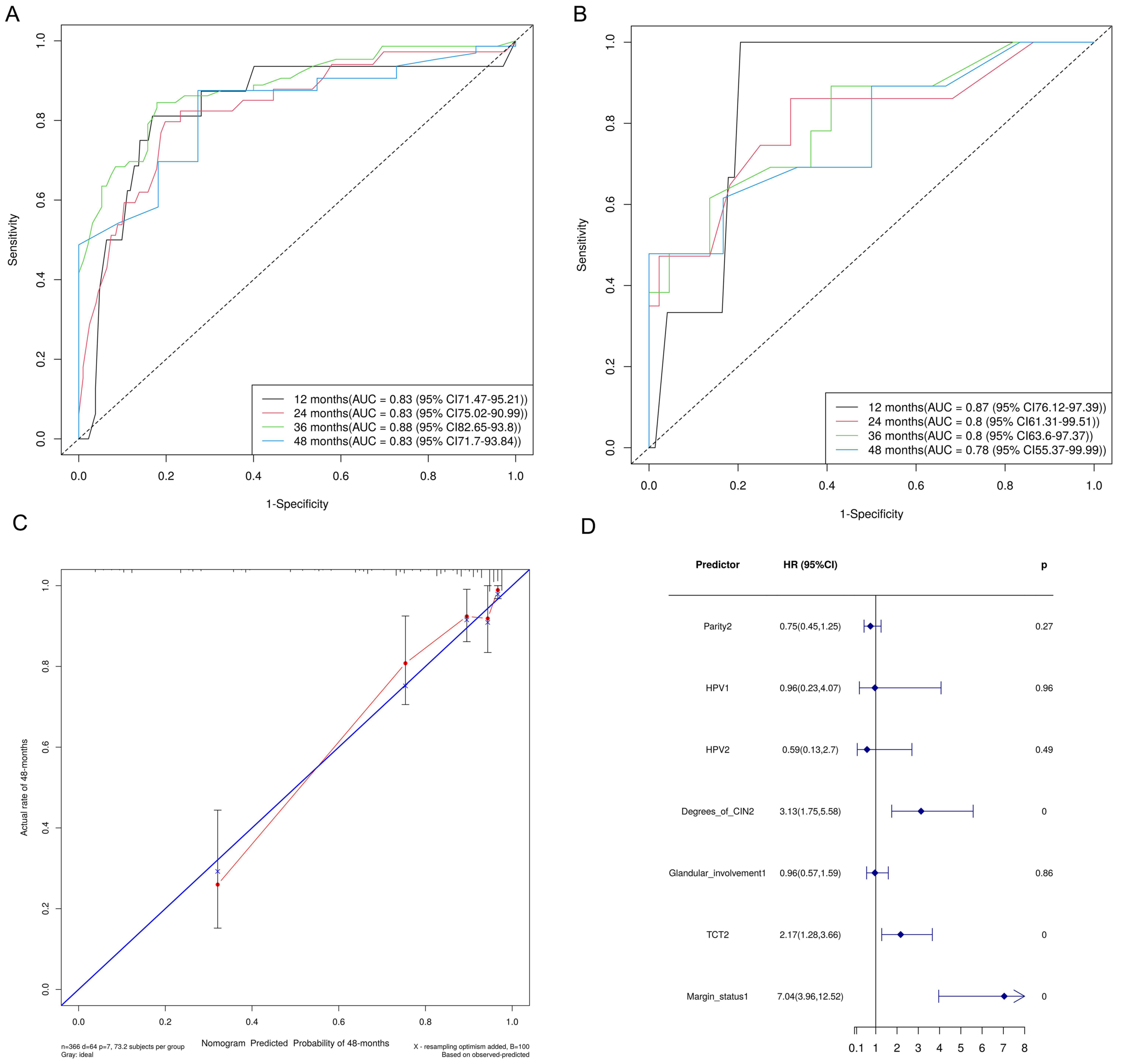


Figure 3 Predictive performance and validation of the Cox proportional hazards model for recurrent or residual disease post-conization in premenopausal women with high-grade cervical intraepithelial neoplasia. Predictive performance of the Cox proportional hazards model in the training (**A**) and testing (**B**) sets across various follow-up periods. (**C**) A calibration curve of the Cox proportional hazards model in the testing set at 48 months. (**D**) A forest plot of the multivariate Cox regression analysis presenting the hazard ratios (HRs) with 95% confidence intervals (CIs) for the various predictors of recurrent or residual disease.

Table 3 The Models´performance in the Training Set and Test Set

Model	AUC				C-Index
	1-Year	2-Year	3-Year	4-Year	
Training set					
RSF-model	0.886	0.862	0.901	0.839	0.802
Cox-model	0.830	0.830	0.880	0.830	0.830
Test set					
RSF-model	0.858	0.790	0.779	0.767	0.802
Cox-model	0.870	0.800	0.800	0.780	0.846

regression model showed lower performance, with an AUC value of 0.830 in most years except for the third year (AUC = 0.880). The superior prediction accuracy and stability of the RSF model underline its effectiveness in handling complex datasets. In the testing set, the two models exhibited similar performances, with the Cox model slightly outperforming the RSF model in the 1 and 2-year predictions (AUC: 0.870 and 0.800 vs 0.858 and 0.790, respectively). This result suggests that the Cox model may be more effective in smaller or specific data subsets. Although the Cox model offered slight advantages over the RSF model in the test scenarios, the RSF model's strong performance in the training set and ability to manage complex interactions emphasize its broader application benefits.

SHAP Value Analysis of the Random Survival Forest Model

This study utilized the RSF model to predict the risk of residual disease and recurrence in premenopausal women with high-grade CIN, with SHAP values being employed to quantify the contribution of each predictor. As illustrated in Figure 4A, "margin status" was the most significant predictor, followed by "degree of CIN", "HPV status", "TCT results", and "parity". Although "glandular involvement" had a relatively lower influence, it still significantly influenced the model's output. The mean SHAP values in Figure 4A represent the average impact magnitude of these predictors, while Figure 4B shows the distribution of these values across the model predictions, with the color scale denoting the effect of the feature value. The predominant positive contributions from "margin status" underscore its crucial role in evaluating the risk of residual disease and recurrence.

Survival Analysis and Risk Stratification

Kaplan–Meier curves were used to distinguish between the high- and low-risk groups identified by the RSF model. Compared to the low-risk group, the high-risk group exhibited a notably reduced survival probability over 60 months ($p < 0.0001$) in the training set, and this trend was similarly observed in the testing set (Figure 5A and B). The tabulated data accompanying each curve present the "number at risk" at various time points, thereby confirming the model's effectiveness in risk stratification and validating the appropriateness of the sample size for this type of analysis.

Model Presentation

The finalized RSF model is accessible through an interactive web application designed for easy replication and validation by peers. This prediction tool can be accessed at http://www.xsmartanalysis.com/model/list/predict/model/html?mid=15006andsymbol=3171PHACMq5272793Ud2, with the general model interface depicted in Figure 6.

Discussion

In this study, we developed a predictive model for residual and recurrent high-grade CIN in premenopausal women following LEEP treatment by applying the RSF algorithm to their clinical and pathological data. The RSF model exhibited better calibration and discrimination capacity in predicting residual and recurrent CIN than traditional Cox regression models. Furthermore, visual analysis of the RSF model highlighted margin status as the most significant

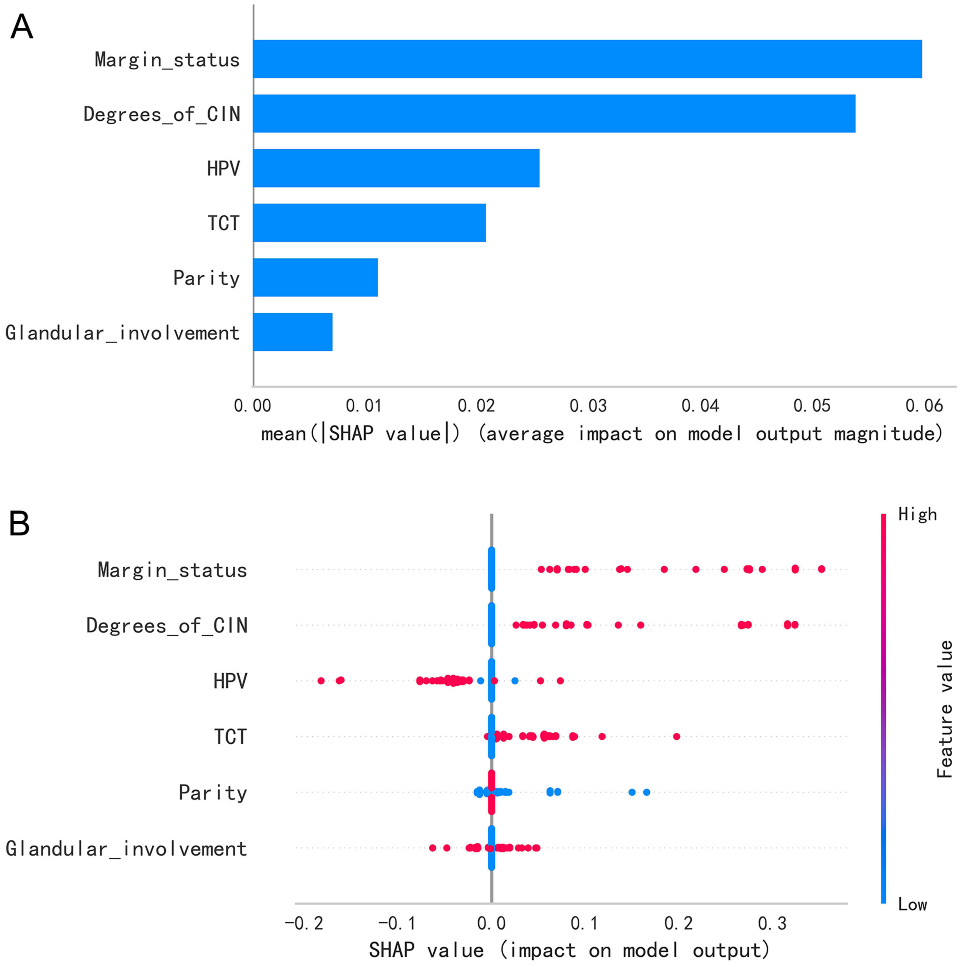


Figure 4 SHAP value analysis of predictive factors in the random forest survival analysis of high-grade cervical intraepithelial neoplasia (CIN) after conization. (**A**) A graph of the mean SHAP values showing the average impact of each predictor on the model output. (**B**) A bee swarm plot displaying the individual SHAP values for each predictor across all data points, with color coding according to the feature value.

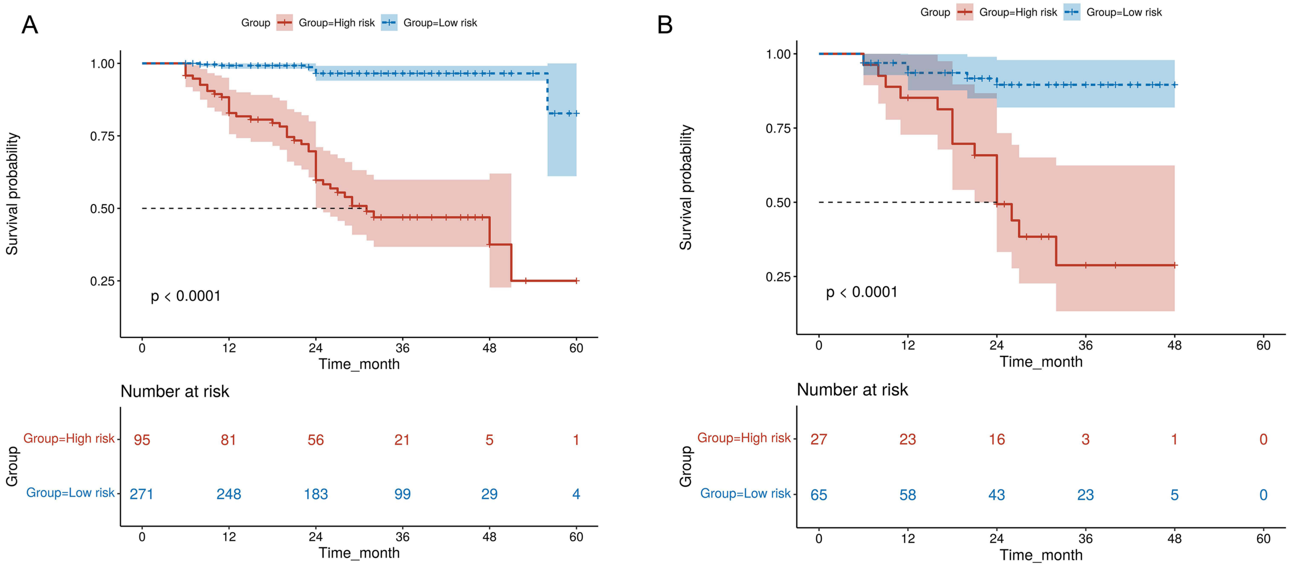
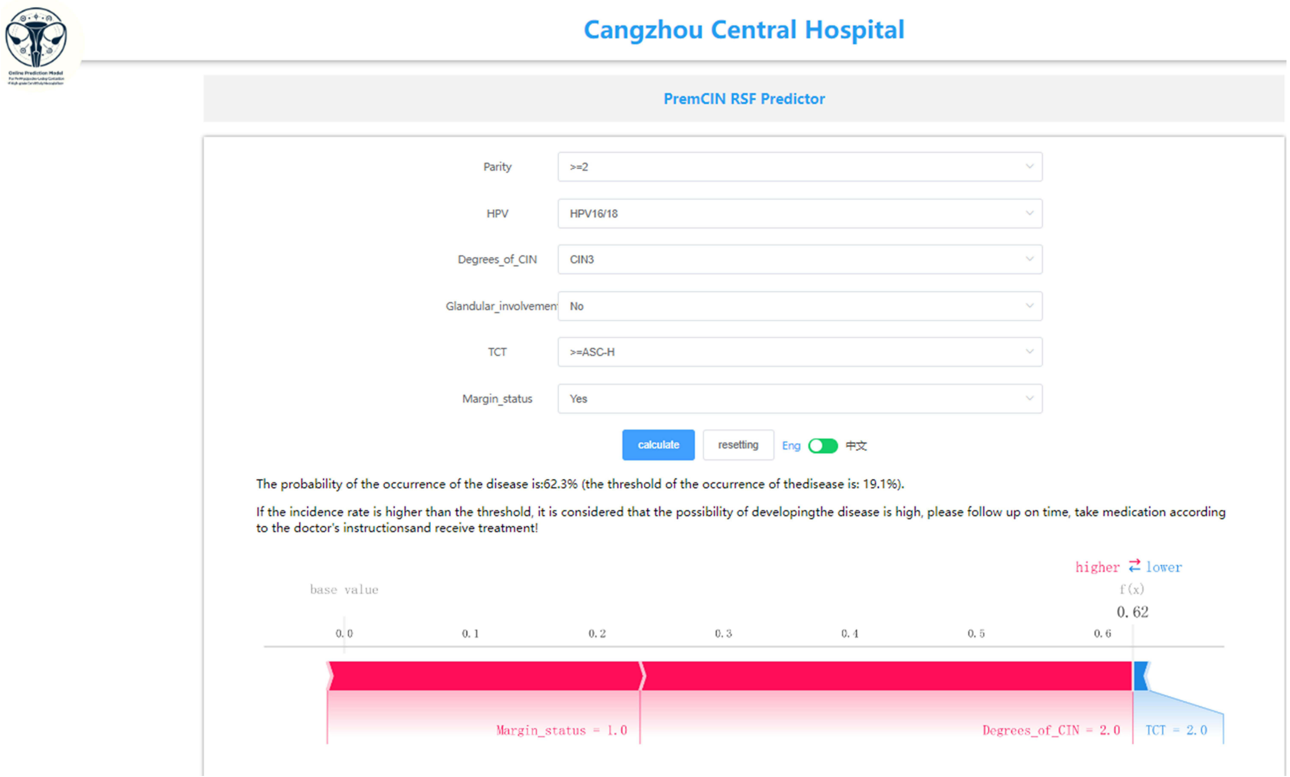


Figure 5 Kaplan–Meier curves for residual and recurrence risk after LEEP for high-grade CIN in premenopausal women. High- and low-risk groups are distinguished over 60 months in the training (**A**) and testing (**B**) sets, along with the indication of the "number at risk" at different intervals and statistical significance.



Cangzhou Central Hospital

PremCIN RSF Predictor

Parity: >=2

HPV: HPV16/18

Degrees_of_CIN: CIN3

Glandular_involvemen: No

TCT: >=ASC-H

Margin_status: Yes

The probability of the occurrence of the disease is:62.3% (the threshold of the occurrence of thedisease is: 19.1%).

If the incidence rate is higher than the threshold, it is considered that the possibility of developingthe disease is high, please follow up on time, take medication according to the doctor's instructionsand receive treatment!

Figure 6 The user interface of the web-based calculator for predicting the risk of residual/recurrent high-grade CIN after conization treatment in premenopausal women.

predictive factor, followed by the degree of CIN, HPV status, and TCT results. The use of the RSF model for risk stratification and individual risk prediction showed great potential in clinical settings, as well as in improving decision-making for personalized treatment approaches.

The Cox regression model has been widely employed in survival analysis and prognosis prediction in various studies, including those on cervical cancer.[25] Prior research has also utilized Cox regression models to assess the risk of residual and recurrent CIN after conization treatment. For example, Bogani et al[15] developed a nomogram that incorporated notable risk factors such as CIN3 diagnosis, high-risk HPV, and positive endocervical margins to predict the persistence or recurrence of cervical dysplasia. Similarly, Ding et al[26] found that high-risk HPV infection, positive surgical margins, and smoking history were critical factors for high-grade CIN recurrence after LEEP, emphasizing the significance of HPV-based surveillance and personalized patient management. In our study, the Cox model showed strong predictive accuracy, with AUC values ranging from 0.83 to 0.88 in the training and testing datasets. However, the model's assumption of a linear relationship between covariates and risk, as well as its reliance on proportional hazards, may limit its applicability in more complex clinical scenarios.[27]

RSF is a machine learning algorithm introduced in 2008 that can effectively manage high-dimensional data and intricate variable interactions, thus overcoming the limitations of the Cox regression model.[28] Our analysis revealed that the RSF model outperformed the Cox model in predicting residual and recurrent high-grade CIN in the specific population of premenopausal women. Moreover, the RSF model consistently exhibited high predictive accuracy, with AUC values ranging from 0.767 to 0.901 in the training and testing sets. Additionally, the RSF model demonstrated excellent prediction stability, with lower OOB error rates and higher CPSs over time. Although the two predictive models showed similar performances in the testing set, the Cox regression model performed slightly better in the 1- and 2-year predictions. Nevertheless, the RSF model's ability to handle complex data and predict long-term outcomes underscored its potential advantages.

Our study aimed to construct a predictive model for residual and recurrent high-grade CIN after conization in premenopausal women. Factors such as surgical margins, HPV infection, lesion size, and immune status are pivotal in evaluating the risk of recurrence and residual disease in premenopausal and post-menopausal women.[10,29–32] However, the relative impact and predictive value of these factors may vary between these two patient groups due to the hormonal, anatomical, and physiological changes associated with menopause.[16,33] Our study applied four algorithms to identify the key predictors of residual and recurrent high-grade CIN in premenopausal women following LEEP. The algorithm results identified six critical variables: margin status, CIN grade, glandular involvement, parity, TCT results, and HPV status.

Further investigation based on the SHAP value analysis quantified the influence of each predictor on the model's predictions. Margin status was determined as the most critical predictor, as demonstrated by a strong correlation between positive margins and a higher risk of residual and recurrent high-grade CIN. This observation was consistent with previous studies that reported positive margins as a reliable indicator of future CIN after conization.[10,34]

In addition to margin status, the severity of CIN and TCT results were also found to play a significant role as predictors. Higher CIN grades, positive TCT results, and the presence of high-risk HPV types were all linked to increased risk for residual and recurrent disease. Other factors such as parity and glandular involvement also notably influenced the risk levels, consistent with previous studies that have highlighted their association with residual and recurrent high-grade CIN following LEEP procedures.[9,10,29,35,36] The ability of the RSF model to incorporate these complex interactions may have contributed to its superior performance over traditional Cox regression models, underscoring its value in clinical prognostic assessments. This finding further emphasizes the usefulness of the RSF model in providing enhanced predictive accuracy and customizing post-operative management strategies for women undergoing LEEP.

Our RSF model effectively stratified the patients into high- and low-risk groups, showing significant differences in survival probabilities over 60 months. This stratification is crucial for identifying patients requiring intensive monitoring or aggressive treatment, such as those in the high-risk category. In contrast to traditional nomograms that predict survival at a specific point without considering individual risk factors, the RSF model offers a more flexible and intuitive approach. This model improves prognostic accuracy by utilizing local SHAP plots to visualize the impact of various risk factors on survival outcomes. Additionally, the RSF model was used to develop an online prediction tool that simplifies patient-specific risk calculations, thereby promoting personalized treatment planning and enhancing care management.

Although this study employs the innovative RSF model and advanced feature selection algorithms, it has certain limitations that should be acknowledged. The retrospective and single-center design as well as the relatively short follow-up period may limit the generalizability and robustness of our study findings. Furthermore, the lack of external validation and the exclusion of variables such as sexual behavior, smoking history, alcohol use, and HPV vaccination status could have potentially undermined the accuracy of the risk predictions. Recent studies underscore the strong association between high-risk HPV infection and various lower genital tract lesions, highlighting the potential therapeutic role of HPV vaccination in reducing recurrence rates of HPV-related lesions following conization in patients with high-grade cervical dysplasia.[37–39] This growing body of evidence suggests that HPV vaccination not only serves as a preventive measure but also reinforces the importance of including vaccination status in future predictive models to enhance patient care. All these unexamined factors are recognized influencers of high-grade CIN outcomes and may have exerted substantial effects on the predictive performance of the study model.

Conclusions

The RSF model constructed in this study is a promising machine learning tool for predicting residual and recurrent high-grade CIN in premenopausal women after LEEP, offering better accuracy and stability than traditional predictive models. Our study findings highlight the potential of personalized management approaches in this patient population, which may contribute to the early prevention and treatment of cervical cancer. Nonetheless, future research should involve larger populations to validate these results and consider integrating other treatment options and preventive measures to enhance the clinical relevance and effectiveness of our study.

Data Sharing Statement

Some or all of the datasets generated and/or analyzed in the current study are not publicly available, but are available on reasonable request by the relevant authors.

Consent to Participate

Informed consent was waived due to the retrospective nature of this study, in line with Article 39 of the "Measures for the Ethical Review of Biomedical Research Involving Humans" issued by the National Health Commission of the People's Republic of China (Order No. 11), which states that informed consent may be waived for "retrospective studies that do not affect the rights and interests of the subjects". This waiver is also consistent with international guidelines, including: The Council for International Organizations of Medical Sciences (CIOMS) Guidelines, specifically Guideline 10 on modifications and waivers of informed consent. The World Health Organization (WHO) Standards and Operational Guidance for Ethics Review of Health-Related Research with Human Participants.

Consent for Publication

All of the authors approved the publication of the article.

Ethics

This study was conducted in accordance with the ethical standards set forth in the 1964 Declaration of Helsinki and its subsequent amendments. This study was approved by the Ethics Committee of Cangzhou Central Hospital (No. 2021-054-02).

Acknowledgments

The authors thank all the doctors and participants involved in the study. This work is supported by the the Extreme Smart Analysis (https://www.xsmartanalysis.com/). Thanks are also extended to Bullet Edits Company for professional language improvement.

Funding

This study was supported by the Key Research project, Hebei Medical Science Research Project (20220350).

Disclosure

The authors declare no competing interests.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2021;71(3):209–249. doi:10.3322/caac.21660
2. Mitra A, Tzafetas M, Lyons D, et al. Cervical intraepithelial neoplasia: screening and management. *Br J Hosp Med*. 2016;77(8):C118–123. doi:10.12968/hmed.2016.77.8.C118
3. Zhu M, He Y, Baak JP, et al. Factors that influence persistence or recurrence of high-grade squamous intraepithelial lesion with positive margins after the loop electrosurgical excision procedure: a retrospective study. *BMC Cancer*. 2015:15744. doi:10.1186/s12885-015-1748-1
4. Yang M, Du J, Lu H, et al. Global trends and age-specific incidence and mortality of cervical cancer from 1990 to 2019: an international comparative study based on the global burden of disease. *BMJ Open*. 2022;12(7):e055470. doi:10.1136/bmjopen-2021-055470
5. Chung SH, Franceschi S, Lambert PF. Estrogen and ERalpha: culprits in cervical cancer? *Trends Endocrinol Metab*. 2010;21(8):504–511. doi:10.1016/j.tem.2010.03.005
6. Zappacosta R, Ianieri MM, Tinelli A, et al. Detection of residual/recurrent cervical disease after successful LEEP conization: the possible role of mRNA-HPV test. *Curr Pharm Des*. 2013;19(8):1450–1457.
7. Melnikow J, McGahan C, Sawaya GF, Ehlen T, Coldman A. Cervical intraepithelial neoplasia outcomes after treatment: long-term follow-up from the British Columbia cohort study. *J Natl Cancer Inst*. 2009;101(10):721–728. doi:10.1093/jnci/djp089
8. Bentivegna E, Maulard A, Pautier P, et al. Fertility results and pregnancy outcomes after conservative treatment of cervical cancer: a systematic review of the literature. *Fertil Steril*. 2016;106(5):1195–1211. doi:10.1016/j.fertnstert.2016.06.032
9. Ikeda M, Mikami M, Yasaka M, et al. Association of menopause, aging and treatment procedures with positive margins after therapeutic cervical conization for CIN 3: a retrospective study of 8856 patients by the Japan society of obstetrics and gynecology. *J Gynecol Oncol*. 2021;32(5):e68. doi:10.3802/jgo.2021.32.e68
10. Andersson S, Megyessi D, Belkic K, et al. Age, margin status, high-risk human papillomavirus and cytology independently predict recurrent high-grade cervical intraepithelial neoplasia up to 6 years after treatment. *Oncol Lett*. 2021;22(3):684. doi:10.3892/ol.2021.12945
11. Wong AS, Li WH, Cheung TH. Predictive factors for residual disease in hysterectomy specimens after conization in early-stage cervical cancer. *Eur J Obstet Gynecol Reprod Biol*. 2016;19921–19926. doi:10.1016/j.ejogrb.2016.01.020
12. Alukal AT, Rema P, Suchetha S, et al. Evaluation of factors affecting margin positivity and persistent disease after leep for cervical intraepithelial neoplasia. *J Obstet Gynaecol India*. 2021;71(4):411–416. doi:10.1007/s13224-021-01450-9
13. Christensen E. Multivariate survival analysis using Cox's regression model. *Hepatol*. 1987;7(6):1346–1358. doi:10.1002/hep.1840070628

14. Bogani G, Tagliabue E, Ferla S, et al. Nomogram-based prediction of cervical dysplasia persistence/recurrence. *Eur J Cancer Prev*. 2019;28 (5):435–440. doi:10.1097/CEJ.0000000000000475
15. Bogani G, Lalli L, Sopracordevole F, et al. Development of a nomogram predicting the risk of persistence/recurrence of cervical dysplasia. *Vaccines*. 2022;10(4):579. doi:10.3390/vaccines10040579
16. Alder S, Megyessi D, Sundstrom K, et al. Incomplete excision of cervical intraepithelial neoplasia as a predictor of the risk of recurrent disease-a 16-year follow-up study. *Am J Obstet Gynecol*. 2020;222(2):e171–172e112. doi:10.1016/j.ajog.2019.08.042
17. Ng'andu NH. An empirical comparison of statistical tests for assessing the proportional hazards assumption of Cox's model. *Stat Med*. 1997;16 (6):611–626. doi:10.1002/(sici)1097-0258(19970330)16:6<611::aid-sim437>3.0.co;2-t
18. Qiu X, Gao J, Yang J, et al. A comparison study of machine learning (Random Survival Forest) and classic statistic (Cox Proportional Hazards) for predicting progression in high-grade glioma after proton and carbon ion radiotherapy. *Front Oncol*. 2020:10551420. doi:10.3389/fonc.2020.551420
19. Wang Y, Deng Y, Tan Y, et al. A comparison of random survival forest and Cox regression for prediction of mortality in patients with hemorrhagic stroke. *BMC Med Inform Decis Mak*. 2023;23(1):215. doi:10.1186/s12911-023-02293-2
20. Ge Y, Liu Y, Cheng Y, Liu Y. Predictors of recurrence in patients with high-grade cervical intraepithelial neoplasia after cervical conization. *Med*. 2021;100(27):e26359. doi:10.1097/MD.0000000000026359
21. Ballout N, Etievant L, Viallon V. On the use of cross-validation for the calibration of the adaptive lasso. *Biom J*. 2023;65(5):e2200047. doi:10.1002/bimj.202200047
22. Kursa MB, Rudnicki WR. Feature selection with boruta package. *J Stat Softw*. 2010;36(11):1–13. doi:10.18637/jss.v036.i11
23. Ding X, Yang F, Ma F. An efficient model selection for linear discriminant function-based recursive feature elimination. *J Biomed Inform*. 2022;129104070. doi:10.1016/j.jbi.2022.104070
24. Urbanowicz RJ, Meeker M, La Cava W, Olson RS, Moore JH. Relief-based feature selection: introduction and review. *J Biomed Inform*. 2018;85189–85203. doi:10.1016/j.jbi.2018.07.014
25. Matsuo K, Purushotham S, Jiang B, et al. Survival outcome prediction in cervical cancer: cox models vs deep-learning model. *Am J Obstet Gynecol*. 2019;220(4):e381–381e314. doi:10.1016/j.ajog.2018.12.030
26. Ding T, Li L, Duan R, et al. Risk factors analysis of recurrent disease after treatment with a loop electrosurgical excision procedure for high-grade cervical intraepithelial neoplasia. *Int J Gynaecol Obstet*. 2023;160(2):538–547. doi:10.1002/ijgo.14340
27. Hess KR. Assessing time-by-covariate interactions in proportional hazards regression models using cubic spline functions. *Stat Med*. 1994;13 (10):1045–1062. doi:10.1002/sim.4780131007
28. Baralou V, Kalpourtzi N, Touloumi G. Individual risk prediction: comparing random forests with Cox proportional-hazards model by a simulation study. *Biom J*. 2023;65(6):e2100380. doi:10.1002/bimj.202100380
29. Chen J-Y, Wang Z-L, Wang Z-Y, Yang X-S. The risk factors of residual lesions and recurrence of the high-grade cervical intraepithelial lesions (HSIL) patients with positive-margin after conization. *Med*. 2018;97(41):e12792. doi:10.1097/MD.0000000000012792
30. Wang X, Xu J, Gao Y, Qu P. Necessity for subsequent surgery in women of child-bearing age with positive margins after conization. *BMC Womens Health*. 2021;21(1):191. doi:10.1186/s12905-021-01329-x
31. Costa S, De Simone P, Venturoli S, et al. Factors predicting human papillomavirus clearance in cervical intraepithelial neoplasia lesions treated by conization. *Gynecol Oncol*. 2003;90(2):358–365. doi:10.1016/s0090-8258(03)00268-3
32. Fan A, Wang C, Han C, et al. Factors affecting residual/recurrent cervical intraepithelial neoplasia after cervical conization with negative margins. *J Med Virol*. 2018;90(9):1541–1548. doi:10.1002/jmv.25208
33. Sun X, Lei H, Xie X, et al. Risk factors for residual disease in hysterectomy specimens after conization in post-menopausal patients with cervical intraepithelial neoplasia grade 3. *Int J Gen Med*. 2020:131067–131074. doi:10.2147/IJGM.S280576
34. Giannini A, Di Donato V, Sopracordevole F, et al. Outcomes of High-Grade cervical dysplasia with positive margins and HPV persistence after cervical conization. *Vaccines*. 2023;11(3):698. doi:10.3390/vaccines11030698
35. Zhao J, Liu X, Gao J, et al. Factors associated with lesion recurrence following cervical conization. *Altern Ther Health Med*. 2023;29(6):50–55.
36. Abdulaziz AMA, You X, Liu L, et al. Management of high-grade squamous intraepithelial lesion patients with positive margin after LEEP conization: a retrospective study. *Med*. 2021;100(20):e26030. doi:10.1097/MD.0000000000026030
37. Zhang J, Liu G, Cui X, Yu H, Wang D. Human papillomavirus genotypes and the risk factors associated with multicentric intraepithelial lesions of the lower genital tract: a retrospective study. *BMC Infect Dis*. 2021;21(1):554. doi:10.1186/s12879-021-06234-0
38. Bogani G, Sopracordevole F, Ciavattini A, et al. HPV-related lesions after hysterectomy for high-grade cervical intraepithelial neoplasia and early-stage cervical cancer: a focus on the potential role of vaccination. *Tumori*. 2024;110(2):139–145. doi:10.1177/03008916231208344
39. Bogani G, Raspagliesi F, Sopracordevole F, et al. Assessing the long-term role of vaccination against HPV after loop electrosurgical excision procedure (LEEP): a Propensity-Score matched comparison. *Vaccines* 2020;8(4):717. doi:10.3390/vaccines8040717