ORIGINAL RESEARCH

Risk Assessment of Pulmonary Metastasis for Cervical Cancer Patients by Ensemble Learning Models: A Large Population Based Real-World Study

Menglin Zhu[1,*]
Bo Wang [2,*]
Tiejun Wang [3]
Yilin Chen[1,4]
Du He[1,5]

[1]Department of Anesthesiology, Hubei Minzu University Affiliated Enshi Clinical Medical School, The Central Hospital of Enshi Tujia and Miao Autonomous Prefecture, Enshi, Hubei, 445000, People's Republic of China; [2]National Clinical Research Center for Obstetrical and Gynecological Diseases; Key Laboratory of Cancer Invasion and Metastasis, Ministry of Education; Department of Obstetrics and Gynecology, Tongji Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, 430030, People's Republic of China; [3]Department of Oncology, Hubei Cancer Hospital, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, People's Republic of China; [4]Department of Pulmonary and Critical Care Medicine, Hubei Minzu University Affiliated Enshi Clinical Medical School, The Central Hospital of Enshi Tujia and Miao Autonomous Prefecture, Enshi, Hubei, 445000, People's Republic of China; [5]Department of Oncology, Hubei Minzu University Affiliated Enshi Clinical Medical School, The Central Hospital of Enshi Tujia and Miao Autonomous Prefecture, Enshi, Hubei, 445000, People's Republic of China

*These authors contributed equally to this work

Correspondence: Du He; Yilin Chen
Email hdu1234@163.com;
chenyl891128@163.com

Objective: Pulmonary metastasis (PM) is an independent risk factor affecting the prognosis of cervical patients, but it still lacks a prediction. This study aimed to develop machine learning-based predictive models for PM.

Methods: A total of 22,766 patients diagnosed with or without PM from the Surveillance, Epidemiology, and End Results (SEER) database were enrolled in this study. The cohort was randomly split into a train set (70%) and a validation set (30%). In addition, 884 Chinese patients from two tertiary medical centers were included as an external validation set. Duplicated and useless candidate variables were excluded, and sixteen variables were included for the machine learning algorithm. We developed five predictive models, including the generalized linear model (GLM), random forest model (RFM), naive Bayesian model (NBM), artificial neural networks model (ANNM), and decision tree model (DTM). The predictive performance of these models was evaluated by the receiver operating characteristic (ROC) curve and calibration curve. The Cox proportional hazard model (CPHM) and competing risk model (CRM) were also included for survival outcome prediction.

Results: Of the patients included in the analysis, 2456 (4.38%) patients were diagnosed with PM. Age, organ-site metastasis (liver, bone, brain), distant lymph metastasis, tumor size, and pathology were the important predictors of PM. The RFM with 9 variables introduced was identified as the best predictive model for PM (AUC = 0.972, 95% CI: 0.958–0.986). The C-index for the CPHM and CRM was 0.626 (95% CI: 0.604–0.648) and 0.611 (95% CI: 0.586–0.636), respectively.

Conclusion: The prediction algorithm derived by machine-learning-based methods shows a robust ability to predict PM. This result suggests that machine learning techniques have the potential to improve the development and validation of predictive modeling in cervical patients with PM.

Keywords: cervical cancer, pulmonary metastasis, machine learning, predictive model, prognosis, SEER database

Introduction

Cervical cancer is the fourth most common cancer worldwide, with more than half a million new cases and 311 365 deaths in 2018.[1,2] According to the US National Comprehensive Cancer Network and the International Federation of Gynecology and Obstetrics (FIGO) guidelines, radical hysterectomy, bilateral pelvic lymphadenectomy, and elective oophorectomy for patients with clinical stage are recommended.[3,4] Women with early-stage cervical cancer who undergo a radical hysterectomy are usually cured, with 5-year disease-free survival (DFS) rates

8713

exceeding 90% in some studies.[5–8] However, delay in the treatment of cancer with advanced stage can have adverse consequences on the outcome, improvements in the clinical stage and effective treatment have led to long-term survival among patients with cervical cancer.[9]

The possibility for cervical cancer patients to develop a subsequent primary cancer becomes a more important consideration for prognosis.[10] An emerging issue in survivors is the occurrence of metastasis formation, which is the major cause of death in most cervical patients with primary cancer. Lymph node metastasis is the main metastatic pathway and the most critical factor for overall survival (OS) and DFS of cervical cancer cases.[11] However, the lymph node status is not involved in the FIGO staging system of uterine cervical cancer, which leads to an absence of important information for prognosis and treatment. Besides, distant site metastasis such as pulmonary, subsequent liver, bone, and brain, are the most common metastasis in cervical cancer.[12] Despite considerable advances in the treatment, patients with distant metastasis still suffer dismal prognosis.[13] Most cervical patients present with asymptomatic, or diagnosed after follow-up visit via imaging examination, which leads to delays in treatment and adverse prognosis.[14] Collectively, we lack standardized estimates of the risk of pulmonary metastasis (PM) for cervical cancer patients.

Machine learning is a branch of computer science and statistics that generates predictive or descriptive models, which has been widely used for data analysis in medicine during the past decade.[15,16] Compared with the conventional linear regression model, the supervised machine learning algorithm has enabled the opportunity to move data science research towards translation for more personalized cancer treatments.[17] This study aimed to develop and validate effective machine learning-based models for the prediction of PM in patients with cervical cancer.

Materials and Methods
Study Population
Patients with cervical cancer were identified from the SEER program of the National Cancer Institute (http:// seer.cancer.gov/). The cohort was composed of patients pathologically confirmed with PM from the SEER database between 2010 and 2016. The accessible data were downloaded SEER*Stat 8.3.6. The primary site code was restricted to the cervix uteri (International Classification of Diseases for Oncology-3/WHO 2008). Since the SEER database began to record metastatic information including pulmonary metastasis from 2010 to 2016. We obtained the information in cervical patients with PM from 1 January 2010 to 31 December 2016. In addition, 884 Chinese patients from two tertiary medical centers (The Central Hospital of Enshi Tujia and Miao Autonomous Prefecture; Hubei Cancer Hospital) were included as an external validation set. This study followed the declaration of Helsinki. Written informed consent was not obtained because the SEER database was publicly available, the data were analyzed retrospectively and anonymously. The flowchart of the patients' selection is listed in Figure 1.

Variables Preparation
A total of 12 clinical characteristics and pathological features were analyzed in this study. However, the variables that can be used for further analyses were including age, race, primary site, cervix tumor size, histological grade, TNM stage (7th ed), number of excised lymph nodes, the outcome of lymph node biopsy, surgery for the cervix or not, location of tumor metastasis, insurance, and marital status. In the SEER database, several methods were introduced to define race. We also redefined race as white, black, and others (American Indian/AK Native, Asian/ Pacific Islander). The grade was defined as unstaged and low, moderate, or high differentiation. The insurance was defined as insured, uninsured, and any Medicaid. The candidate variables, except age, were defined as categorical variables. Overall survival (OS) was defined as the date of diagnosis to the date of death due to any cause or the date of the last follow-up. Cancer-specific survival (CSS) was calculated as the time from the date of diagnosis to the date of death attributed to cervical cancer or the date of the last follow-up.

Construction of Machine Learning-Based Models
All patients were randomly divided into training and testing groups at a ratio of 7 to 3. Five algorithms were applied to predict PM, including the generalized linear model (GLM), random forest model (RFM), naive Bayesian model (NBM), artificial neural networks model (ANNM), and decision tree model (DTM).[17–19] Both the Cox proportional hazard model (CPHM) and competing risk model (CRM) were evaluated in this study.[20,21]
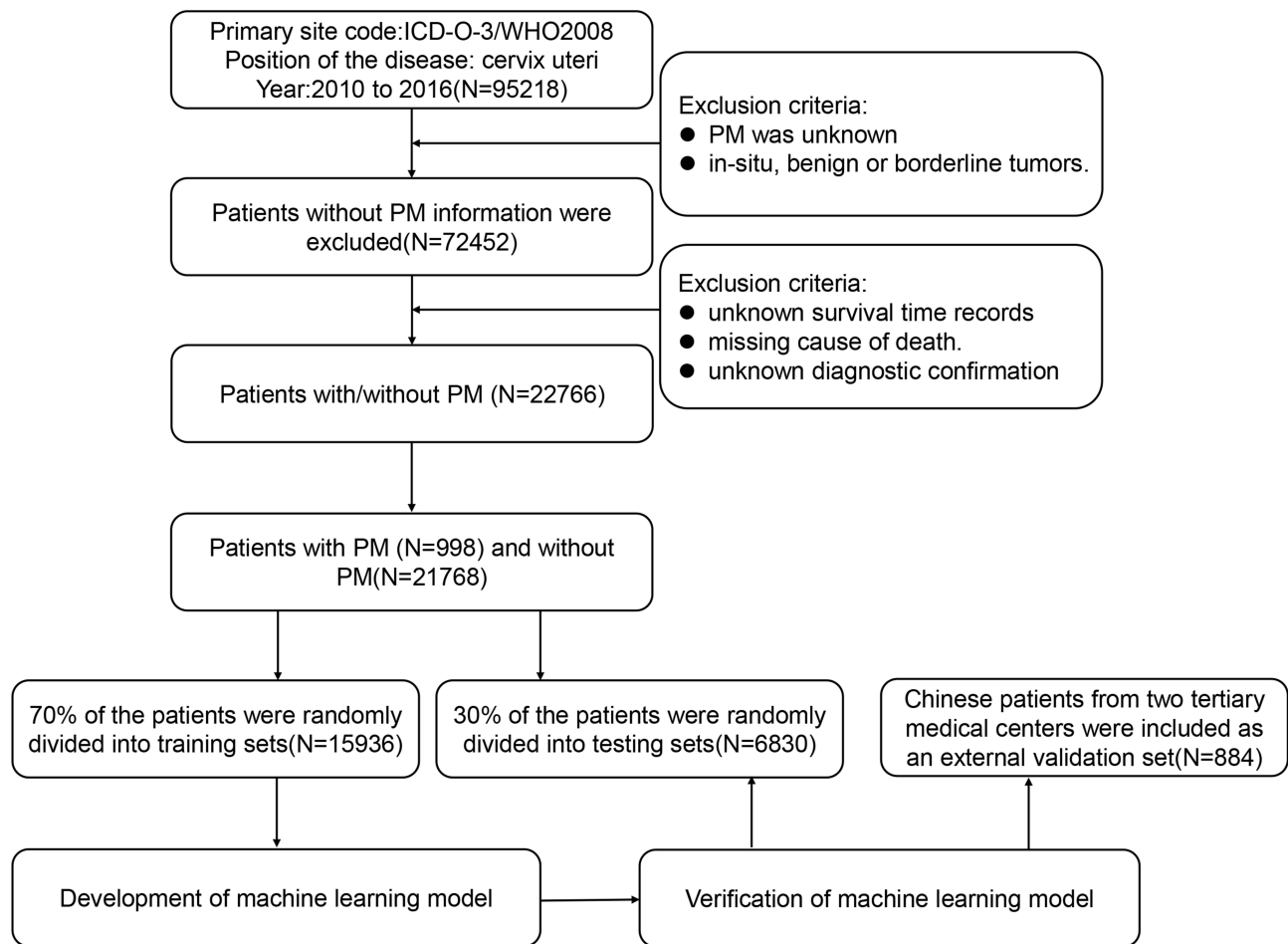
Figure 1 The flow chart of selecting the cervical patients with PM.

Feature Selection and Evaluation of Models

To minimize the negative influence of overfitting in machine learning algorithms, tenfold cross-validation (CV) was repeated ten times to evaluate the accuracy of models. To rank and select meaningful variables, the optimal variables in different algorithms were ranked, and the variables with high ranks were developed for the establishment of prediction models. Differences of models between predicted and actual data were also recorded. We also used the area under the curve (AUC) of the receiver operating characteristic (ROC) curve to evaluate the performances of different models.

Statistical Analysis

For descriptive analysis, continuous variables were presented as the mean with standard deviation or as median with interquartile range (IQR). Categorical variables are presented as numbers (%). The OS, CSS, and risk

assessment of PM were estimated using the nomogram score. The nomograms were added across dependent variables to derive total points, which were converted to predicted probabilities. The predictive performance of the nomogram predictive model was measured by concordance index (C-index) and calibration with 1000 bootstrap samples to decrease the overfit bias. All analysis was performed using the Python programming language (version 3.9.2, Python Software Foundation, https://www.python.org/) and R Project for Statistical Computing (version 4.0.4, http://www.r-project.org/). All P values were two-tailed, and $P<0.05$ was considered statistically significant.

Results
Demographic and Clinicopathological Characteristics

A total of 22,766 cervical cancer patients diagnosed with PM or not were collected in this study. A correlation

matrix was spotted, and sixteen candidate variables correlated with PM were summarized in Supplementary Figure 1. The whole data set was randomly and automatically divided into a training set (N=15936, 70%) and a testing set (N=6830, 30%). Comparing to patients without pulmonary metastasis, the impacts of age, race, higher grade, pathology (adenocarcinoma vs squamous cell carcinoma), location of tumor invasion, TNM stage, lymph biopsy, surgery, regional lymph nodes examination, distant site metastases (liver, brain, and lung), and tumor size (≥5cm) contributed to higher PM incidence. In 884 external data sets, 19 (2.15%) patients were diagnosed with PM. The clinicopathological characteristics of patients with or without pulmonary metastasis were presented in Table 1.

Assessment of Risk Factors for PM in Patients with Cervical Cancer

The GLM, as a traditional linear regression model, was first developed for multivariable analysis, dealing with both category and continuous variables. The best model has then to be identified based on the Akaike Information Criterion (AIC). Candidate variables (AIC=5982.0) such as age at diagnosis, tumor size, grade, T stage, bone metastasis, brain metastasis, and liver metastasis were significantly associated with PM (Supplementary Table 1).

These independently associated risk factors were used to form a PM risk estimation nomogram (Figure 2A). The resulting model was internally validated using the bootstrap validation method. The AUC (0.81, 95% CI:0.79–0.82) Brier score (3.5, 95% CI:3.3–3.7) measures the accuracy of probabilistic predictions (Figure 2B). The predicted accuracy of the nomogram was also validated via the risk histogram (Figure 2C). Besides, the random allocation was used to assess the RFM, the importance of variables could be reflected by the mean decreased Gini (MDG) index. As indicated in Figure 3A, sixteen variables were ordered according to the MDG index. Consistent with the variables screened by GLM, TNM stage, pathology, tumor size, liver metastasis, bone metastasis, brain metastasis, and distant lymph metastasis was more significant for PM prediction (Supplementary Table 2). The relationship of variables changes between the prediction error and the number of decision trees were shown in Figure 3B. A total of 6830 testing samples were adopted to assess the efficacy of the prediction model, and the results revealed that RFM performed better than GLM (Figure 3C).

Comparison Between Machine Learning-Based Models

The predictive performance of the five models was shown in Supplementary Table 3. The best performance was observed in the RFM (AUC =0.97, 95% CI:0.96–0.99), which performed similarly to DT (AUC =0.96, 95% CI:0.95–0.96), ANN (AUC=0.94, 95% CI:0.82–0.99), and NBM (AUC=0.92, 95% CI:0.91–0.93). Moreover, all models performed significantly better than conventional methods (GLM). Consistent with the results of training set, the external data set also confirmed that the RFM model was optimal (AUC =0.95, 95% CI:0.46–1.44).

Survival Outcomes and Prognostic Factors for Cervical Cancer Patients with PM

The CPHM and CRM were also included for survival outcome prediction. The multivariate Cox analysis showed that age at diagnosis, surgery, race, bone metastasis, brain metastasis, liver metastasis, and distant lymph metastasis significantly correlated with OS (Supplementary Table 4). Similarly, the age at diagnosis, surgery, bone metastasis, brain metastasis, liver metastasis, and distant lymph metastasis were also correlated with CSS (Supplementary Table 5). Based on this, independent prognostic factors were used to establish the OS and CSS nomograms (Figures 4 and 5). The C-index of the OS nomogram was 0.69 and the concordance index of the CSS nomogram was 0.63. In addition, validation showed consistent agreement between the risk of survival outcome and the actual estimation by the testing cohort.

Discussion

For patients with cervical cancer, 4.16% to 7.7% of patients develop PM.[1,22] Although PM from primary cervical carcinoma is rare, the number of metastatic nodules, pulmonary metastasectomy, and the disease-free interval between the primary gynecologic procedure is tightly associated with survival outcome.[14,23] Nowadays, preclinical studies have shown that PD-1 blockade can inhibit tumor growth and even reduce metastasis, which may bring a glimmer of hope for the prognosis of PM in patients with cervical cancer.[24]

Table 1 Baseline Demographic and Clinical Characteristics of Included Patients Diagnosed with and without PM

Variables	Level	Total (N=22766)	Yes (N=998)	No (N=21768)	P-value
Age, y (median [IQR])		49.00 [39.00, 61.00]	58.00 [49.00, 68.]	49.00 [39.00, 61.00]	<0.001
Race (%)	White	16,957 (74.5)	731 (73.2)	16,226 (74.5)	<0.001
	Black	3153 (13.8)	172 (17.2)	2981 (13.7)	
	Other	2403 (10.6)	95 (9.5)	2308 (10.6)	
	Unknown	253 (1.1)	0 (0.0)	253 (1.2)	
Year (%)	2010	3220 (14.1)	113 (11.3)	3107 (14.3)	0.038
	2011	3193 (14.0)	126 (12.6)	3067 (14.1)	
	2012	3242 (14.2)	152 (15.2)	3090 (14.2)	
	2013	3079 (13.5)	130 (13.0)	2949 (13.5)	
	2014	3309 (14.5)	150 (15.0)	3159 (14.5)	
	2015	3364 (14.8)	173 (17.3)	3191 (14.7)	
	2016	3359 (14.8)	154 (15.4)	3205 (14.7)	
Site* (%)	Cervix uteri	17,682 (77.7)	870 (87.2)	16,812 (77.2)	<0.001
	Endocervix	4321 (19.0)	108 (10.8)	4213 (19.4)	
	Exocervix	411 (1.8)	10 (1.0)	401 (1.8)	
	OLC	352 (1.5)	10 (1.0)	342 (1.6)	
Grade (%)	Grade I	2536 (11.1)	18 (1.8)	2518 (11.6)	<0.001
	Grade II	7029 (30.9)	172 (17.2)	6857 (31.5)	
	Grade III	6427 (28.2)	415 (41.6)	6012 (27.6)	
	Grade IV	542 (2.4)	46 (4.6)	496 (2.3)	
	Unknown	6232 (27.4)	347 (34.8)	5885 (27.0)	
Pathology (%)	ADC	4276 (18.8)	141 (14.1)	4135 (19.0)	<0.001
	SCC	14552 (63.9)	573 (57.4)	13,979 (64.2)	
	Others	3938 (17.3)	284 (28.5)	3654 (16.8)	
SEER sit$^{£}$ (%)	Localized	8788 (38.6)	0 (0.0)	8788 (40.4)	<0.001
	Regional	7421 (32.6)	0 (0.0)	7421 (34.1)	
	Distant	2822 (12.4)	844 (84.6)	1978 (9.1)	
	Unknown	3735 (16.4)	154 (15.4)	3581 (16.5)	
T stage^{δ} (%)	T0	13 (0.1)	2 (0.2)	11 (0.1)	<0.001
	T1	10,462 (46.0)	90 (9.0)	10,372 (47.6)	
	T2	4242 (18.6)	143 (14.3)	4099 (18.8)	
	T3	3045 (13.4)	323 (32.4)	2722 (12.5)	
	T4	724 (3.2)	112 (11.2)	612 (2.8)	
	TX	767 (3.4)	159 (15.9)	608 (2.8)	
	Unknown	3513 (15.4)	169 (16.9)	3344 (15.4)	
N stage (%)	N0	13,676 (60.1)	223 (22.3)	13,453 (61.8)	<0.001
	N1	4834 (21.2)	472 (47.3)	4362 (20.0)	
	NX	743 (3.3)	134 (13.4)	609 (2.8)	
	Unknown	3513 (15.4)	169 (16.9)	3344 (15.4)	
M stage (%)	M0	16,632 (73.1)	0 (0.0)	16,632 (76.4)	<0.001
	M1	2621 (11.5)	829 (83.1)	1792 (8.2)	
	Unknown	3513 (15.4)	169 (16.9)	3344 (15.4)	
Lym biopsy (%)	<3	466 (2.0)	9 (0.9)	457 (2.1)	<0.001
	≥4	7264 (31.9)	27 (2.7)	7237 (33.2)	
	Unknown	15,036 (66.0)	962 (96.4)	14,074 (64.7)	

(*Continued*)

Table 1 (Continued).

| Variables	Level	Total (N=22766)	Yes (N=998)	No (N=21768)	P-value
Surgery (%)	Yes No Unknown	12,582 (55.3) 10,030 (44.1) 154 (0.7)	85 (8.5) 909 (91.1) 4 (0.4)	12,497 (57.4) 9121 (41.9) 150 (0.7)	<0.001
Lym examination (%)	Negative Positive Unknown	14,281 (62.7) 8161 (35.8) 324 (1.4)	909 (91.1) 60 (6.0) 29 (2.9)	13,372 (61.4) 8101 (37.2) 295 (1.4)	<0.001
Bone met (%)	Yes No Unknown	537 (2.4) 22,185 (97.4) 44 (0.2)	223 (22.3) 750 (75.2) 25 (2.5)	314 (1.4) 21,435 (98.5) 19 (0.1)	<0.001
Brain met (%)	Yes No Unknown	85 (0.4) 22,640 (99.4) 41 (0.2)	48 (4.8) 925 (92.7) 25 (2.5)	37 (0.2) 21,715 (99.8) 16 (0.1)	<0.001
Liver met (%)	Yes No Unknown	486 (2.1) 22,245 (97.7) 35 (0.2)	244 (24.4) 734 (73.5) 20 (2.0)	242 (1.1) 21,511 (98.8) 15 (0.1)	<0.001
Lym distant (%)	Yes No Unknown	269 (1.2) 3076 (13.5) 19,421 (85.3)	73 (7.3) 77 (7.7) 848 (85.0)	196 (0.9) 2999 (13.8) 18,573 (85.3)	<0.001
Tumor size, cm (%)	<5 ≥5 Unknown	13,256 (58.2) 455 (2.0) 9055 (39.8)	426 (42.7) 1 (0.1) 571 (57.2)	12,830 (58.9) 454 (2.1) 8484 (39.0)	<0.001
Insurance (%)	Medicaid Insured Uninsured Unknown	6881 (30.2) 13,794 (60.6) 1412 (6.2) 679 (3.0)	355 (35.6) 553 (55.4) 65 (6.5) 25 (2.5)	6526 (30.0) 13,241 (60.8) 1347 (6.2) 654 (3.0)	0.001
Marital status (%)	Married Unmarried Unknown	14,345 (63.0) 6984 (30.7) 1437 (6.3)	641 (64.2) 307 (30.8) 50 (5.0)	13,704 (63.0) 6677 (30.7) 1387 (6.4)	0.217

Notes: *According to the primary site labeled. §According to the American Joint Committee on Cancer(AJCC), 6th. £According to the SEER historic stage (1973–2015). **Abbreviations:** IQR, interquartile range; OLC, overlapping lesion of cervix uteri (cervix uteri, endocervix, exocervix equivalent FIGO I, overlapping lesion of cervix uteri equivalent FIGO II); ADC, adenocarcinoma; SCC, squamous cell carcinoma; Lym biopsy, lymph node biopsy; Lym exam, lymph node examination; Bone met, bone metastasis; Brain met, brain metastasis; Liver met, liver metastasis; Lym distant, distant lymphatic metastasis.

Owing to this lack of a specific and practical predictive method, the development of a predictive model that incorporates factors associated with PM based on clinicopathological characteristics data becomes desirable. This population-based study explored the relationship between cervical cancer patients with PM, survival outcomes, and risk factors, which is essential for designing effective treatment strategies. To our knowledge, this study is the first to investigate the risk factors and prognostic factors in cervical cancer patients associated with PM based on machine learning algorithms.

In this study, using machine learning algorithms, we developed five models to predict PM in cervical cancer

patients, the ROC analysis was used to evaluate the predictive performance and clinical values of these models, respectively. Based on the supervised learning method, most predictive models maintained high AUCs (ranging from 0.919 to 0.972), all of these models performed better than conventional logistic regression analysis (AUC was 0.805) in the prediction of PM. Thus, reliable predictive models for PM in cervical cancer patients were constructed via machine learning, an accurate estimation of PM presence can help surgeons choose appropriate treatment procedures based on risk-benefit assessment.

Previous studies revealed that senior age, nonsquamous type, advanced stage, lymph nodes metastases,

A

B

C
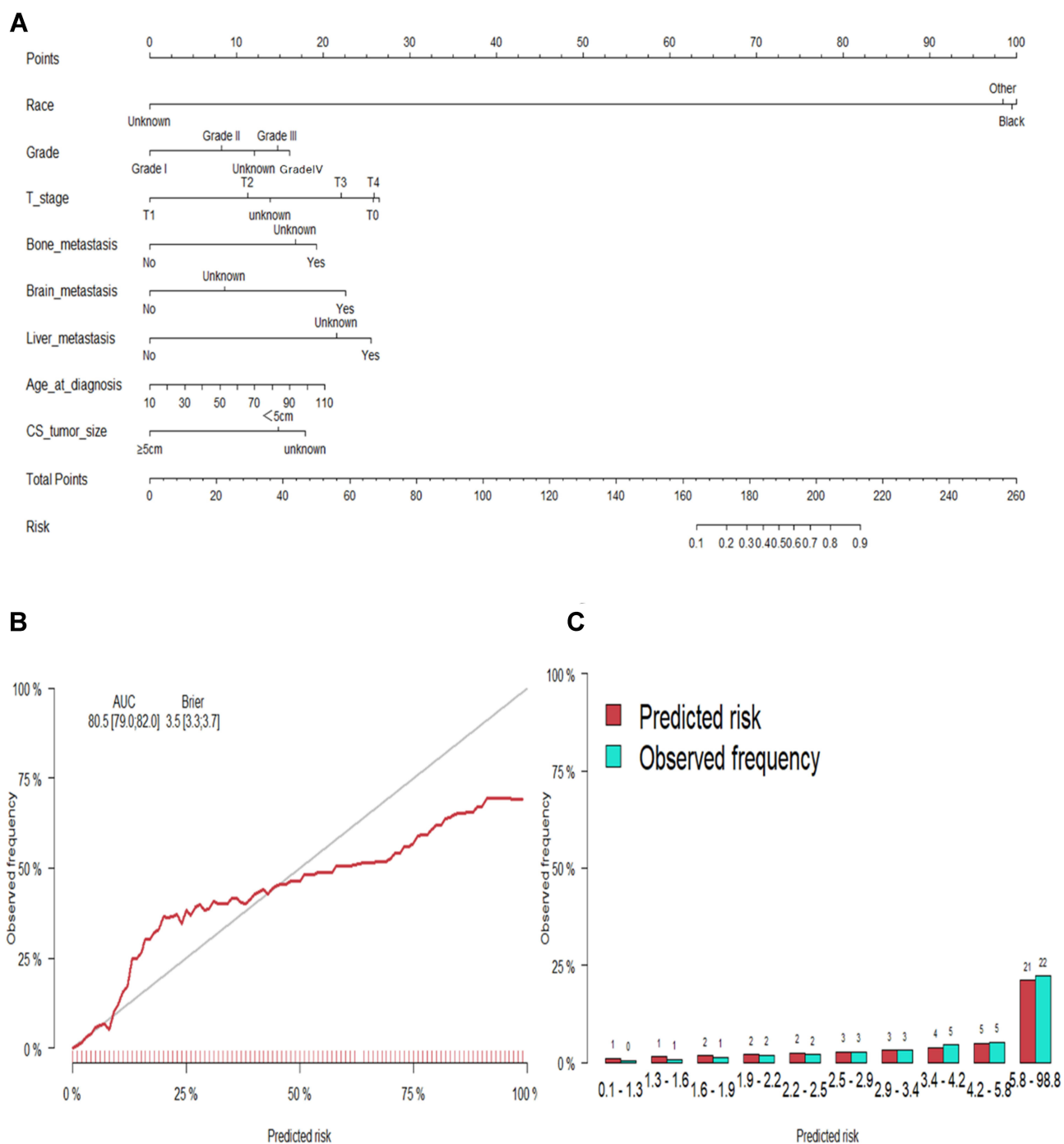
Figure 2 Generalized linear model. (**A**) Nomograms conveying the results of the candidate factors for predicting micrometastasis of lymph nodes. (**B**) Calibration curves for internal validation of the nomogram. (**C**) Predicted risk histogram comparing predicted risk of the nomogram with the observed frequency.

and poor differentiation were associated with an increased risk for PM.[1,14,23,25] In this study, other factors that have been recognized to be important for PM formation were not included in the predictive model, such as organ-site metastasis, TNM stage. Of the currently available prediction tools, the predictive performance indicated that the random forest algorithm was most effective than other models in distinguishing the possibility of PM. Unlike other established risk assessment methods, the random forest algorithm can achieve higher accuracy in the disease prediction by using bootstrap aggregation and randomization of predictors.[26,27] Interestingly, using machine learning algorithms, we found that the age, tumor size, histological grade, organ-site metastasis, pathologic
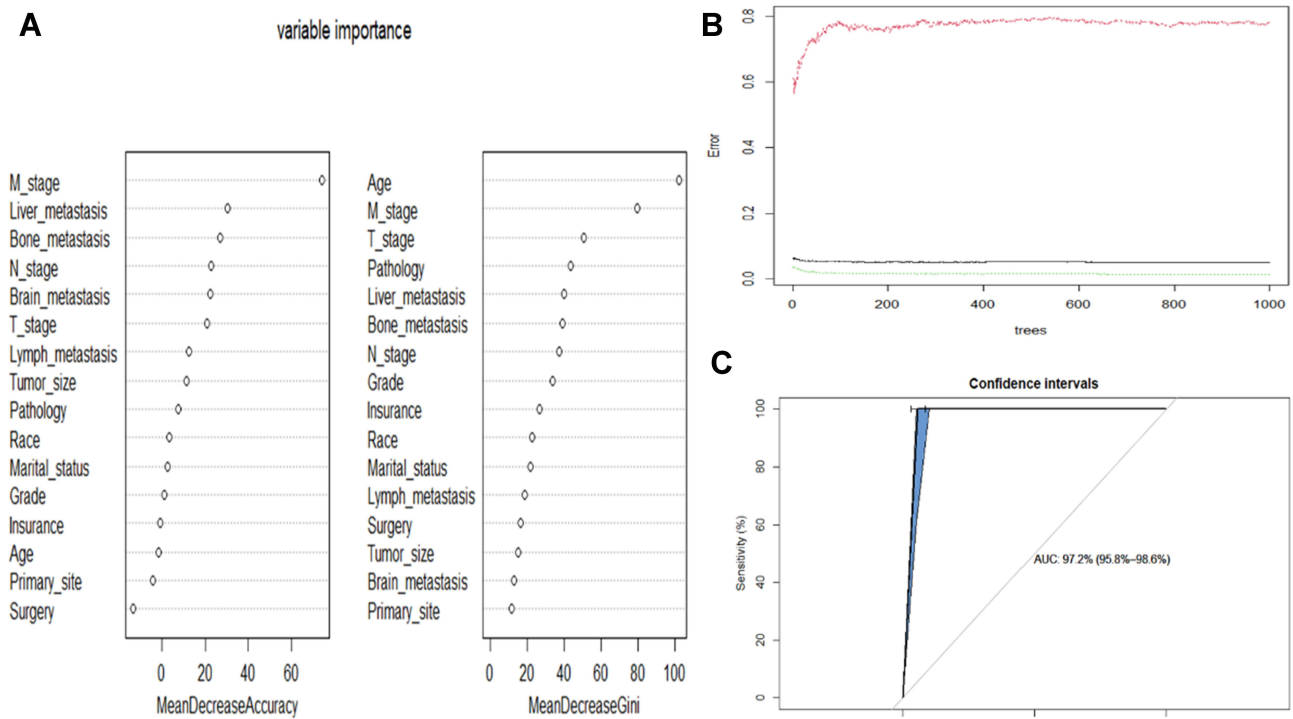
Figure 3 Random forest model. (**A**) The candidate factors associated with micrometastasis of lymph nodes were ordered according to the mean decreased Gini index. (**B**) Relationship of dynamic changes between the prediction error and the number of decision trees. (**C**) Performance of the prediction model with increasing numbers of features in the ROC curve.
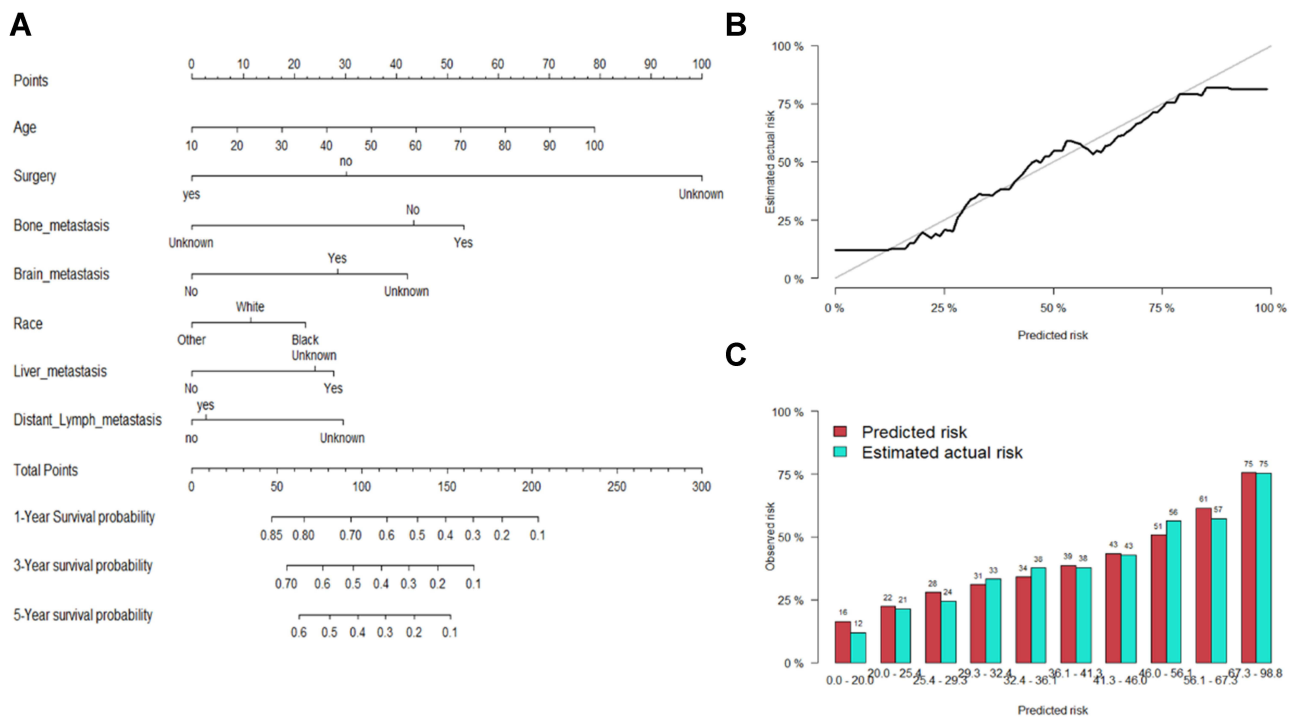
Figure 4 Cox proportional hazard model. (**A**) Nomograms for 1-, 3-, and 5-year overall survival (OS) prediction. (**B**) Calibration curves for internal validation of the nomogram. (**C**) Predicted risk histogram comparing predicted risk of the nomogram with the observed frequency.
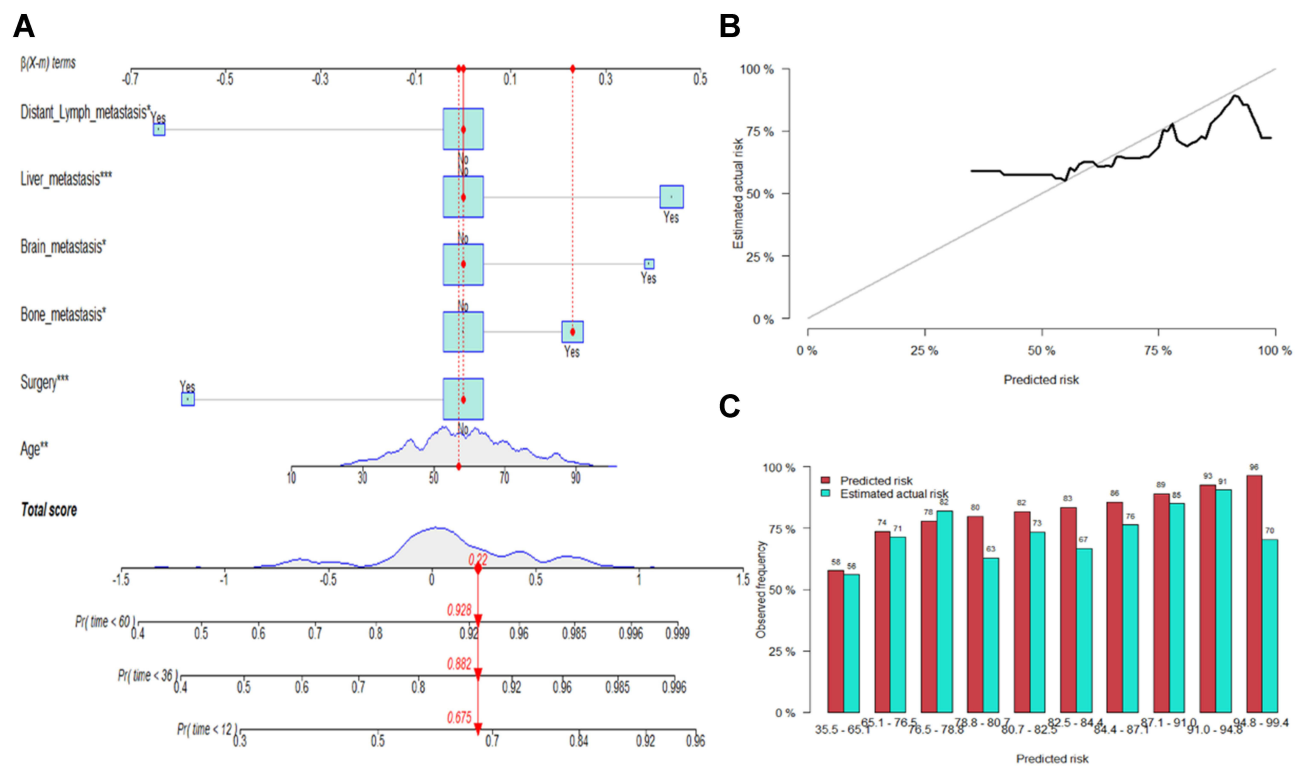
Figure 5 Competitive risk model. (**A**) Nomogram predicting CSS at 1-, 3- and 5-year using the competitive risk model. According to the Nomogram score, patient No.31 has a cumulative risk of 0.675, 0.882, and 0.928 at 1-, 3- and 5-year, respectively. (**B**) Calibration curves for internal validation of the nomogram. (**C**) Predicted risk histogram comparing predicted risk of the nomogram with the observed frequency. *P<0.05; ***P<0.001.

types, and TNM stage were indicated to be the most contributive risk factors of PM, consistent with the results of linear regression analysis. However, in this study, the aforementioned factors were included, and correlations within those clinicopathological characteristics were considered, which accounted for a better result than multivariable logistic regression.

Until now, the Cox proportional hazard model has been the most widely applied to predict survival outcomes. Due to the rarity of PM in cervical cancer patients, limited studies as to prognosis have been launched before.[12,28,29] In the present study, we identified several factors, including older age, surgery, race, organ site metastasis (bone, brain, liver), and distant lymph metastasis, were identified as prognostic factors for cervical cancer patients with PM at diagnosis. The median survival after PM was eighteen months with 2- and 5-year survival rates of 37.7% and 7.5%, respectively.[1] Traditionally, several widely recognized staging systems are available for evaluating the prognosis of cervical cancer patients, such as the Federation International of Gynecology and Obstetrics (FIGO) staging, TNM staging systems, and American Joint

Committee (AJCC) cancer staging system. However, these staging systems and prognostic factors were only applied to cervical cancer patients without PM. In this study, we developed the proposed nomogram, which incorporated seven comprehensive and easily available variables, performed well as supported by the C index values in the training and validation cohorts, respectively. In the PM risk estimation nomogram, age, surgery, organ site metastasis, and distant lymph metastasis have been reported to decrease the possibility of better prognosis in advanced cervical cancer.[1,30–32] Besides, when cervical malignancy metastasizes to the pulmonary, surgery is a good choice because it provides a survival advantage.[33,34] Collectively, based on these survival predictions, the nomogram might serve as a tool to select patients for evaluating the efficacy of pulmonary resection in patients with early cervical cancer and different risks of PM.

This large-scale study revealed the clinical characteristics, risk, and prognostic factors for cervical cancer patients with PM. However, there were also some limitations in our study. First, this analysis was based on data from the SEER database, as mentioned above, information

on active surgical resection of the pulmonary lesion(s), and non-operative therapy could not be obtained, which was recommended to improve the prognosis; Second, although the nomogram achieved definite predictive accuracy of PM, however, the estimation of prognosis for the cervical cancer patient with PM remained relative inefficiency, extra laboratory parameters, which might play crucial roles in determining the prognosis; Third, this was a retrospective study that could not completely avoid data selection and measurement biases, it is necessary to validate the results from more prospective studies or multi-center studies in the future.

Conclusion

The deep learning signature could be used for risk stratification in cervical cancer patients with or without PM, and the random forest model provided an optimal estimation of PM risk in cervical cancer patients. The feature selection approach identified that age, surgery, advanced stage, organ site metastasis, lymph nodes metastases, and poor differentiation were associated with an increased risk for PM, some of which also correlated with the survival outcome.

Data Sharing Statement

Our research project was conducted ethically in accordance with the World Medical Association Declaration of Helsinki. Permission for collection and analysis of the data has been authorized by the Surveillance, Epidemiology, and End Results Program (SEER ID:15914-Nov2019; https://seer.cancer.gov/). All detailed data included in the study are available upon request by contacting the corresponding author.

The raw data of this study are derived from the SEER database

Ethics and Patient Consent Statement

All procedures performed in studies involving human participants were following the ethical standards of the institutional and national research committee and with the Helsinki declaration. This study was approved by the Research Ethics Commission of Tongji Medical College, Huazhong University of Science and Technology (TJ-IRB20210631) with waived informed consent by the Ethics Commission mentioned above.

References

1. Li H, Wu X, Cheng X. Advances in diagnosis and treatment of metastatic cervical cancer. *J Gynecol Oncol*. 2016;27(4):e43. doi:10.3802/jgo.2016.27.e43
2. Stelzle D, Tanaka LF, Lee KK, et al. Estimates of the global burden of cervical cancer associated with HIV. *Lancet Glob Health*. 2021;9(2):e161–e169. doi:10.1016/S2214-109X(20)30459-9
3. Koh WJ, Abu-Rustum NR, Bean S, et al. Cervical cancer, version 3.2019, NCCN clinical practice guidelines in oncology. *JNCCN*. 2019;17(1):64–84. doi:10.6004/jnccn.2019.0001
4. Bhatla N, Denny L. FIGO cancer report 2018. *Int J Gynaecol Obstet*. 2018;143(Suppl 2):2–3. doi:10.1002/ijgo.12608
5. Sedlis A, Bundy BN, Rotman MZ, Lentz SS, Muderspach LI, Zaino RJ. A randomized trial of pelvic radiation therapy versus no further therapy in selected patients with stage IB carcinoma of the cervix after radical hysterectomy and pelvic lymphadenectomy: a Gynecologic Oncology Group Study. *Gynecol Oncol*. 1999;73(2):177–183. doi:10.1006/gyno.1999.5387
6. Peters WA, Liu PY, Barrett RJ, et al. Concurrent chemotherapy and pelvic radiation therapy compared with pelvic radiation therapy alone as adjuvant therapy after radical surgery in high-risk early-stage cancer of the cervix. *J Clin Oncol*. 2000;18(8):1606–1613. doi:10.1200/JCO.2000.18.8.1606
7. Ramirez PT, Frumovitz M, Pareja R, et al. Minimally invasive versus abdominal radical hysterectomy for cervical cancer. *N Engl J Med*. 2018;379(20):1895–1904. doi:10.1056/NEJMoa1806395
8. Nitecki R, Ramirez PT, Frumovitz M, et al. Survival after minimally invasive vs open radical hysterectomy for early-stage cervical cancer: a systematic review and meta-analysis. *JAMA Oncol*. 2020;6(7):1019–1027. doi:10.1001/jamaoncol.2020.1694
9. Hanna TP, King WD, Thibodeau S, et al. Mortality due to cancer treatment delay: systematic review and meta-analysis. *BMJ*. 2020;371:m4087. doi:10.1136/bmj.m4087

10. Mariotto AB, Rowland JH, Ries LA, Scoppa S, Feuer EJ. Multiple cancer prevalence: a growing challenge in long-term survivorship. *Cancer Epidemiol Biomark Prev.* 2007;16(3):566–571. doi:10.1158/1055-9965.EPI-06-0782

11. Huang BX, Fang F. Progress in the study of lymph node metastasis in early-stage cervical cancer. *Curr Med Sci.* 2018;38(4):567–574. doi:10.1007/s11596-018-1915-0

12. Zhang S, Wang X, Li Z, Wang W, Wang L. Score for the overall survival probability of patients with first-diagnosed distantly metastatic cervical cancer: a novel nomogram-based risk assessment system. *Front Oncol.* 2019;9:1106. doi:10.3389/fonc.2019.01106

13. van Meir H, Kenter GG, Burggraaf J, et al. The need for improvement of the treatment of advanced and metastatic cervical cancer, the rationale for combined chemo-immunotherapy. *Anticancer Agents Med Chem.* 2014;14(2):190–203. doi:10.2174/18715206113136660372

14. Ki EY, Lee KH, Park JS, Hur SY. A clinicopathological review of pulmonary metastasis from uterine cervical cancer. *Cancer Res Treat.* 2016;48(1):266–272. doi:10.4143/crt.2014.206

15. Radakovich N, Nagy M, Nazha A. Machine learning in haematological malignancies. *Lancet Haematol.* 2020;7(7):e541–e50. doi:10.1016/S2352-3026(20)30121-6

16. Thrall JH, Li X, Li Q, et al. Artificial intelligence and machine learning in radiology: opportunities, challenges, pitfalls, and criteria for success. *JACR.* 2018;15(3 Pt B):504–508. doi:10.1016/j.jacr.2017.12.026

17. Badillo S, Banfai B, Birzele F, et al. An introduction to machine learning. *Clin Pharmacol Ther.* 2020;107(4):871–885. doi:10.1002/cpt.1796

18. Ngiam KY, Khor IW. Big data and machine learning algorithms for health-care delivery. *Lancet Oncol.* 2019;20(5):e262–e273. doi:10.1016/S1470-2045(19)30149-4

19. Cuocolo R, Caruso M, Perillo T, Ugga L, Petretta M. Machine learning in oncology: a clinical appraisal. *Cancer Lett.* 2020;481:55–62. doi:10.1016/j.canlet.2020.03.032

20. Austin PC, Fine JP. Practical recommendations for reporting Fine-Gray model analyses for competing risk data. *Stat Med.* 2017;36(27):4391–4400. doi:10.1002/sim.7501

21. Jiang H, Fine JP. Survival analysis. *Methods Mol Biol.* 2007;404:303–318.

22. Ishikawa M, Nakayama K, Rahman MT, et al. [A case of stage IVb cervical carcinoma in which survival was prolonged by two different chemotherapies and CCRT]. *Gan to Kagaku Ryoho.* 2012;39(3):451–455. Japanese.

23. Clavero JM, Deschamps C, Cassivi SD, et al. Gynecologic cancers: factors affecting survival after pulmonary metastasectomy. *Ann Thorac Surg.* 2006;81(6):2004–2007. doi:10.1016/j.athoracsur.2006.01.068

24. McDermott J, Jimeno A. Pembrolizumab: PD-1 inhibition as a therapeutic strategy in cancer. *Drugs Today.* 2015;51(1):7–20. doi:10.1358/dot.2015.51.1.2250387

25. Chen X, Chen L, Zhu H, Tao J. Risk factors and prognostic predictors for Cervical Cancer patients with lung metastasis. *J Cancer.* 2020;11(20):5880–5889. doi:10.7150/jca.46258

26. Rigatti SJ. Random forest. *J Insurance Med.* 2017;47(1):31–39.

27. Su X, Xu Y, Tan Z, et al. Prediction for cardiovascular diseases based on laboratory data: an analysis of random forest model. *J Clin Lab Anal.* 2020;34(9):e23421. doi:10.1002/jcla.23421

28. Yamamoto K, Yoshikawa H, Shiromizu K, et al. Pulmonary metastasectomy for uterine cervical cancer: a multivariate analysis. *Ann Thorac Surg.* 2004;77(4):1179–1182. doi:10.1016/j.athoracsur.2003.06.023

29. Carlson V, Delclos L, Fletcher GH. Distant metastases in squamous-cell carcinoma of the uterine cervix. *Radiology.* 1967;88(5):961–966. doi:10.1148/88.5.961

30. Frumovitz M, Querleu D, Gil-Moreno A, et al. Lymphadenectomy in locally advanced cervical cancer study (LiLACS): phase III clinical trial comparing surgical with radiologic staging in patients with stages IB2-IVA cervical cancer. *J Minim Invasive Gynecol.* 2014;21(1):3–8. doi:10.1016/j.jmig.2013.07.007

31. Nartthanarung A, Thanapprapasr K, Udomsubpayakul U, Thanapprapasr D. Age and survival of cervical cancer patients with bone metastasis. *APJCP.* 2014;15(19):8401–8404. doi:10.7314/apjcp.2014.15.19.8401

32. Sperduto PW, Berkey B, Gaspar LE, Mehta M, Curran W. A new prognostic index and comparison to three other indices for patients with brain metastases: an analysis of 1960 patients in the RTOG database. *Int J Radiat Oncol Biol Phys.* 2008;70(2):510–514. doi:10.1016/j.ijrobp.2007.06.074

33. Shu T, Bai P, Zhang R, Li S. [Clinical analysis and prognostic factors in 106 patients with stage Ia-IIb cervical cancer with pulmonary metastasis]. *Zhonghua Zhong Liu Za Zhi.* 2014;36(9):703–707. Chinese.

34. González Casaurrán G, Simón Adiego C, Peñalver Pascual R, Moreno Mata N, Lozano Barriuso M, González Aragoneses F. Surgery of female genital tract tumour lung metastases. *Arch Bronconeumol.* 2011;47(3):134–137. doi:10.1016/S1579-2129(11)70032-5