Received: 3 May 2022 | Revised: 3 July 2022 | Accepted: 9 September 2022

DOI: 10.1111/aogs.14462

ORIGINAL RESEARCH ARTICLE

The application of risk models based on machine learning to predict endometriosis-associated ovarian cancer in patients with endometriosis

Xiaopei Chao[1,2] | Shu Wang[1,2] | Jinghe Lang[1,2] | Jinhua Leng[1,2] | Qingbo Fan[1,2]

[1]Department of Obstetrics and Gynecology, Peking Union Medical College Hospital (PUMCH), Chinese Academy of Medical Sciences (CAMS) & Peking Union Medical College, Beijing, China

[2]National Clinical Research Center for Obstetric & Gynecologic Diseases, Beijing, China

Correspondence
Shu Wang, Department of Obstetrics and Gynecology, Peking Union Medical College Hospital, Shuaifuyuan No. 1, Dongcheng District, Beijing 100730, China.
Email: wangshu219@hotmail.com

Funding information
CAMS Innovation Fund for Medical Sciences, Grant/Award Number: CIFMS 2021-I2M-1-003; National Key R&D Program of China, Grant/Award Number: 2017YFC1001200

Abstract

Introduction: There is currently no satisfactory model for predicting malignant transformation of endometriosis. The aim of this study was to construct and evaluate a risk model incorporating noninvasive clinical parameters to predict endometriosis-associated ovarian cancer (EAOC) in patients with endometriosis.

Material and Methods: We enrolled 6809 patients with endometriosis confirmed by pathology, and randomly allocated them to training (n = 4766) and testing cohorts (n = 2043). The proportion of patients with EAOC in each cohort was similar. We extracted a total of 94 demographic and clinicopathologic features from the medical records using natural language processing. We used a machine learning method – gradient-boosting decision tree – to construct a predictive model for EAOC and to evaluate the accuracy of the model. We also constructed a multivariate logistic regression model inclusive of the EAOC-associated risk factors using a back stepwise procedure. Then we compared the performance of the two risk-predicting models using DeLong's test.

Results: The occurrence of EAOC was 1.84% in this study. The logistic regression model comprised 10 selected features and demonstrated good discrimination in the testing cohort, with an area under the curve (AUC) of 0.891 (95% confidence interval [CI] 0.821–0.960), sensitivity of 88.9%, and specificity of 76.7%. The risk model based on machine learning had an AUC of 0.942 (95% CI 0.914–0.969), sensitivity of 86.8%, and specificity of 86.7%. The machine learning-based risk model performed better than the logistic regression model in DeLong's test (p = 0.036). Furthermore, in a prospective dataset, the machine learning-based risk model had an AUC of 0.8758, a sensitivity of 94.4%, and a specificity of 73.8%.

Conclusions: The machine learning-based risk model was constructed to predict EAOC and had high sensitivity and specificity. This model could be of considerable use in helping reduce medical costs and designing follow-up schedules.

Abbreviations: AUC, area under the ROC curve; DCA, decision curve analysis; EAOC, endometriosis-associated ovarian cancer; GnRHa, gonadotrophin-releasing hormone agonists; NLP, natural language processing; ROC, receiver operating characteristic curve.

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.
© 2022 The Authors. *Acta Obstetricia et Gynecologica Scandinavica* published by John Wiley & Sons Ltd on behalf of Nordic Federation of Societies of Obstetrics and Gynecology (NFOG).

KEYWORDS
endometriosis, machine learning, malignant transformation, ovarian cancer, risk model

1 | INTRODUCTION

The prevalence of endometriosis has been estimated at 5%–10% among women of childbearing age.[1] It has been reported that approximately 1% of patients with endometriosis experience a malignancy, the most common site (80%) of which was in the ovaries,[2] with the presence of associated endometriosis either in the same ovary or elsewhere.[3] Ovarian cancer along with endometriosis is termed endometriosis-associated ovarian cancer (EAOC). Patients with endometriosis-associated clear cell ovarian cancer should also receive retroperitoneal staging and adjuvant therapy.[4,5] Compared with epithelial ovarian cancer, EAOC has a younger age onset, an earlier stage of disease and lower grade at diagnosis, decreased recurrence rates (26.9% vs. 41%), and a better overall 5-year survival (75% vs. 55%).[6,7]

In spite of being a rare disease and having a better prognosis than epithelial ovarian cancer, EAOC decreases patients' survival. Furthermore, the screening for and monitoring of endometriosis malignancy is more likely to be harmful because of the unnecessary medical care and invasive examinations.[8] With more accurate and noninvasive diagnostic methods for EAOC, the overdiagnosis and overtreatment associated with EAOC may be avoided.[9] Thus, an ability to predict the probability of malignancy among patients with endometriosis remains of great significance.

After substantial exploration, endometriosis malignancy-related risk factors have been gradually revealed. Increasingly innovative methods, including biochemical and imaging technical advancements, have been proposed to predict endometriosis malignancy, but they remain less than satisfactory. Therefore, a more practical, noninvasive, and cost-effective risk model is urgently needed. The use of artificial intelligence tools may be another promising method to predict EAOC, and has been widely used in clinical practice.

The objective of this study was to construct and evaluate a risk-predicting model based on machine learning depending on epidemiologic and clinicopathologic features extracted from electronic medical records. This risk model can help gynecologists predict which patients are more likely to experience malignancy transformation of endometriosis and to provide appropriate and timely intervention.

Key message

No practical risk model yet exists to predict malignant transformation of endometriosis. This machine learning-based risk model can help predict the chances of malignant transformation of epidemiologically and clinicopathologically characterized endometriosis.

2 | MATERIAL AND METHODS

2.1 | Study design and patient enrollment

Detailed medical records of patients with pathologically confirmed endometriosis were extracted from patients admitted to our national referral hospital for surgery from January 1, 2015, to September 1, 2019, using natural language processing (NLP). With the logistic regression model as a comparison, the machine learning algorithms were adopted to exploit the most appropriate model for predicting the risk of endometriosis malignancy. The same NLP method was used to extract the prospective dataset for validation from September 1, 2019, to May 30, 2021. The flow chart showing this process is presented in Figure 1.

The inclusion criteria were as follows: (1) patients aged ≥18 years and (2) diagnosed with "endometriosis" and concurrent "ovarian cancer", namely patients with endometriosis confirmed by surgery and/or pathology and concurrent ovarian cancer confirmed by pathology. Patients with repeat cases and those with a diagnosis of endometriosis not confirmed by pathology and/or surgery were excluded.

2.2 | Intervention

A number of criteria are used to define EAOC. Strict histologic conditions are applied in the Sampson and Scott criteria for diagnosis of cancer resulting from endometriosis, which is malignant transformation in the endometriosis glands resulting in EAOC.[10,11] The Van Gorp classification provides broader criteria,[12] with EAOC divided into three categories: (A) ovarian cancers with histologic proof of areas of transition between endometriosis, atypical endometriosis, and endometriosis-associated carcinoma in accordance with Sampson and Scott's definition; (B) ovarian cancers with endometriosis in the same ovary, but without histologic proof of transition; (C) ovarian cancers with endometriosis at any location in the pelvis. We applied Van Gorp's criteria in this study.

2.3 | Data collection and processing

Patients' medical information, including epidemiologic and clinicopathologic data, were extracted using algorithms. NLP techniques were applied to extract information efficiently and accurately, which included Chinese word segmentation, national word identification, part-of-speech tagging, semantic analysis, parsing, and relation extraction. Regular expression (RegEx)

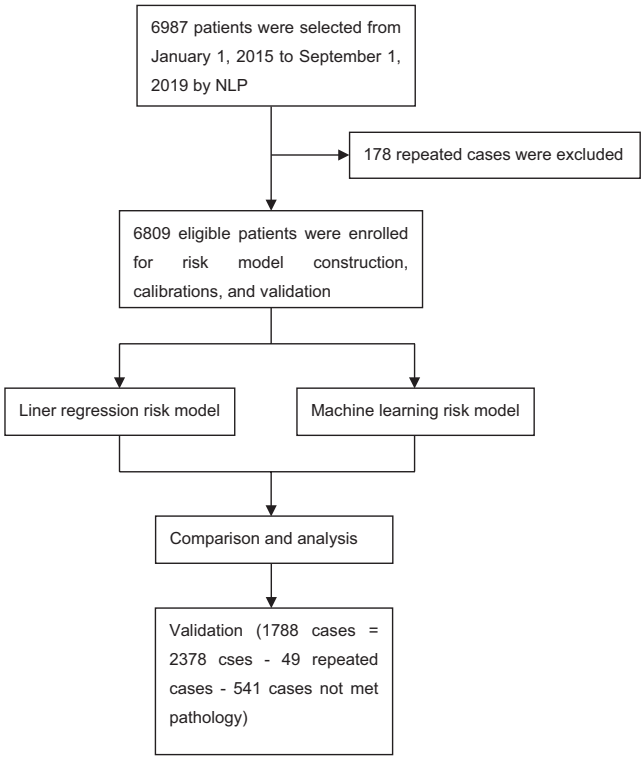

FIGURE 1 Flow chart showing study processes. NLP, natural language processing.

was also employed to match the keywords. The information was extracted in three steps: (1) sentence segmentation and word tokenization of the medical history or examination reports, (2) part-of-speech tagging and normalization of the text processed through step 1, and (3) key information extraction according to the semantic logic between words.[13] A total of 178 repeated cases were excluded from the dataset, and 6809 patients were finally enrolled for the analysis. The patients were divided into those with or without EAOC, and a total of 94 variables were included. Finally, a 6809×95 matrix was generated based on this rule, in which each row represented a patient and each column represented a type of clinical feature.

2.4 | Statistical analyses

Normally distributed continuous variables were described as means ± standard deviations, and non-normally distributed discrete variables were summarized as median (range) and interquartile range. Categorical variables were expressed as counts with percentages. For univariate analysis, independent *t*-tests were conducted as appropriate to compare continuous data between groups if assumptions of normal distribution were confirmed. The Mann–Whitney U test was used for non-normally distributed variables, and the chi-squared test was used for categorical data analysis. *p* values <0.05 were considered statistically significant. All data analyses were performed using R 3.6.2 (R Core Team) and Python 3.7.

2.5 | Machine learning algorithms

Since the data distribution was extremely unbalanced, models were trained and tested on a series of datasets that were split by ratios from 9:1 to 1:9 (starting with 9:1, then 8:2, and so on until 1:9) for trains and tests to validate the robustness of the model. The collected cases were randomly separated into training and testing datasets, and the proportion of cases with or without EAOC were the same in both datasets. While developing the machine learning model, 10-fold cross validation was applied while the training dataset was divided into 10 equivalent parts containing the same number of cases with EAOC, of which nine were used as the training set and the remaining one as the testing set. This process continued until each part was used once for validation.

In the derivation dataset, we used a gradient-boosting decision tree model powered by LightGBM (v3.0.0) to predict patients with endometriosis resulting in EAOC using all the predictor variables. This method generates a sequence of decision trees, and each tree is used to optimize the outcomes that were predicted poorly by the previous trees. The algorithm allows for the development of a prediction model by first creating a single decision tree that best identifies patients at risk for EAOC. The algorithm then creates a subsequent decision tree designed to identify patients who developed EAOC and who were not predicted accurately by the first tree (i.e., the residuals of the first tree). This process continues iteratively, with each additional tree aiming to better predict cases of EAOC missed by the previous trees. The final model, therefore, consists of a series of trees, with the optimal number of leaf nodes in the derivation dataset.

2.6 | The logistic regression risk model

To ensure the comparability of results, the same training and testing datasets were used for the model construction as in the final machine learning algorithms. All covariates with a *p*-value <0.1 were initially enrolled in the multivariate logistic regression analysis. A backward, stepwise procedure was used to select the best combination of potential predictors that were independently associated with the diagnosis of EAOC in the training dataset, and we then evaluated the model with the testing data. The odds ratio (OR) of enrolled variables with 95% confidence interval (CI) were reported. The risk model was assessed by area under the receiver operating characteristic (ROC) curve (AUC) and calibration curve. Changes in the model were also observed when each enrolled variable was removed. The sensitivity and specificity of the final models of the training set and the testing set were both reported.

2.7 | Comparison and validation of the risk models

To compare the performances of the two models, the AUCs were compared using DeLong's test.[14] In addition, the established model

was also validated with the prospective dataset collected in the same hospital using the same NLP method.

2.8 | Ethics statement

This study was approved by the institutional review board from the study center (no. JS-1532) on March 27, 2018. Informed consent was obtained before enrollment. All procedures performed in the study involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Declaration of Helsinki and its later amendments or comparable ethical standards.

3 | RESULTS

3.1 | Patients' characteristics

A total of 6809 patients were finally enrolled for the analysis and risk model construction, 125 (1.84%) of whom were diagnosed with EAOC. Compared with the non-EAOC group, the EAOC group was significantly older at diagnosis (46.7 ± 9.1 vs. 36.2 ± 8.2 years; $p < 0.001$), were older at menopause (49.1 ± 4.3 vs. 47.3 ± 6.2 years; $p = 0.023$), had a higher proportion of patients in menopause ($p < 0.001$), and had less parity ($p < 0.001$). As for the chief complaints, a lower proportion of EAOC patients had dysmenorrhea ($p < 0.001$), dyspareunia ($p = 0.001$), or infertility ($p < 0.001$), but a higher proportion had a pelvic mass ($p < 0.001$). Furthermore, a higher proportion of patients with EAOC had personal history of endometrial carcinoma ($p < 0.001$). A history of gynecologic surgery was common in the EAOC group and included benign ovarian cystectomy ($p = 0.003$), oophorectomy ($p < 0.001$), and salpingectomy ($p = 0.010$). Serum carbohydrate antigen 125 (CA125) and CA199 was higher in those patients with EAOC, and the maximum diameter of the tumor was larger. Patient baseline characteristics are summarized in Table 1.

3.2 | Logistic regression model

3.2.1 | Data analysis

The final logistic regression model included 10 variables (Table 2). Seven variables with OR ≥1 were positively related with the occurrence of EAOC. However, chief complaints of dysmenorrhea, preoperative use of gonadotrophin-releasing hormone agonists (GnRHa), and concurrent leiomyoma or adenomyoma were negatively associated with malignant transformation of endometriosis. This model had an AUC of 0.903 (95% CI 0.857–0.948) with a sensitivity of 89.2% and a specificity of 82.3% in the training dataset (Figure 2A) and an AUC of 0.891 (95% CI 0.821–0.960) with a sensitivity of 88.9% and a specificity of 76.7% in the testing dataset (Figure 2B).

3.2.2 | Model evaluation

To further explore the potency of this model, the calibration of the model was assessed by calibration curve[15] as shown in Figure 2C. The *p*-value of the Hosmer-Lemeshow goodness-of-fit test was 0.4, which was >0.05, whereas the C-index was 0.891. Thus, the calibration curve also verified the prediction value of this model. Changes in the predictive performance were observed when each variable was removed. The results showed that the use of GnRHa played the most important part among all the variables. Thus, this model was further evaluated in those without the use of GnRHa and revealed an AUC of 0.891 (95% CI 0.850–0.931) (Figure 2D), with a sensitivity of 84.3% and a specificity of 83.8%.

3.3 | Machine learning risk model

3.3.1 | Data analysis

It is noteworthy that only 125 of the total 6809 patients were diagnosed with EAOC. Since the data distribution was extremely unbalanced, models were trained and tested with a series of datasets that were split by ratios from 9:1 to 1:9. Even when leaf nodes were set at the minimum of two, the AUC for the most unbalanced datasets (i.e., 1:9) could still reach over 0.79, indicating that the model derived based on this dataset would be robust (Table S1). The top 10 variables were selected as independent predictors of highest significance from the model with the most optimal hyperparameters. The features of significance for each of the variables were reported in Figure 3.

3.3.2 | Model evaluation

The number of nodes was optimized in the derivation dataset to maximize the AUC. The hyperparameters with the highest AUC in the testing dataset, which indicated the performance of the model, were considered as optimal hyperparameters. The AUC with sensitivity and specificity for testing datasets of different hyperparameters are shown in Table S2. The gradient-boosting machine with the most optimal hyperparameters (leaf nodes: 10, training: testing = 7:3) had the largest AUC (0.9417; 95% CI 0.914–0.969) in the testing dataset, with a sensitivity of 86.8% and a specificity of 86.7% (Figure 4).

3.4 | Model comparison

The AUC increased from 0.891 as in the logistic regression model to 0.942 as in the machine learning risk model in the testing dataset (Figure 5). The difference was statistically significant ($p = 0.036$) (Table S3).

TABLE 1 Epidemiologic and clinicopathological characteristics of the patients with endometriosis

Characteristic	Total cases ($n = 6809$)	EAOC ($n = 125$)	Non-EAOC ($n = 6684$)	p value
Age at diagnosis (years)	36.4 ± 8.4	46.7 ± 9.1	36.2 ± 8.2	**<0.001**
Height (cm)	162.9 ± 9.6	162.2 ± 5.0	162.9 ± 9.6	0.421
Weight (kg)	58.9 ± 10.0	60.32 ± 7.7	58.8 ± 10.1	**0.042**
BMI (kg/m^2)	22.9 ± 27.5	22.9 ± 2.7	22.9 ± 27.8	0.990
Age of menarche (years)	13.5 ± 12.7	13.9 ± 1.5	13.5 ± 12.8	0.694
Menopause				**<0.001**
No	6476 (95.1)	80 (64.0)	6396 (95.7)	
Yes	333 (4.9)	45 (36.0)	288 (4.3)	
Age of menopause (years)	47.6 ± 6.0	49.1 ± 4.3	47.3 ± 6.2	**0.023**
Gravidity (times)	1.0 (0–11)	2.0 (0–7)	1 (0–11)	**<0.001**
Parity (times)	1.0 (0–4)	1.0 (0–3)	1.0 (0–4)	**<0.001**
Labor method				**<0.001**
VD	828 (12.2)	30 (24.0)	798 (11.9)	
CS	1367 (20.1)	27 (21.6)	1340 (20.0)	
VD and CS	35 (0.5)	2 (1.6)	33 (0.5)	
Nulliparous	2591 (38.1)	25 (20.0)	2566 (38.4)	
Chief complaints				**<0.001**
Dysmenorrhea	3319 (48.7)	24 (19.2)	2531 (37.9)	**<0.001**
Abdominal pain	1138 (16.7)	30 (24.0)	1108 (16.6)	0.055
Pelvic mass	3571 (52.4)	95 (76.0)	3476 (52.0)	**<0.001**
Infertility	1357 (19.9)	2 (1.6)	1355 (20.3)	**<0.001**
AUB	578 (8.5)	12 (9.6)	566 (8.5)	0.390
Increased menstruation	547 (8.0)	0 (0)	547 (8.2)	**<0.001**
Dyspareunia	447 (6.6)	1 (0.8)	446 (6.7)	**0.001**
Urinary symptoms	445 (6.5)	14 (11.2)	431 (6.4)	0.055
Gastrointestinal symptoms	620 (9.1)	12 (9.6)	608 (9.1)	0.407
Medical complications				0.868
Hypertension	296 (4.4)	15 (12.0)	281 (4.2)	**<0.001**
Diabetes mellitus	128 (1.9)	5 (4.0)	123 (1.8)	0.078
Hepatic disease	258 (3.8)	12 (9.6)	246 (3.7)	**0.001**
Renal disease	46 (0.7)	2 (1.6)	44 (0.7)	0.207
Autoimmune disease	43 (0.6)	1 (0.8)	42 (0.6)	0.550
Thyroid disease	253 (3.7)	8 (6.4)	245 (3.7)	0.109
Personal history of tumor				0.351
Endometrial carcinoma	99 (1.5)	10 (8.0)	89 (1.3)	**<0.001**
Intestinal carcinoma	20 (0.3)	1 (0.8)	19 (0.3)	0.310
Breast cancer	71 (1.0)	3 (2.4)	68 (1.0)	0.141
Thyroid carcinoma	60 (0.9)	2 (1.6)	58 (0.9)	0.302
Hematological tumor	5 (0.1)	0 (0)	5 (0.1)	0.911
Family history of tumor				0.285
Ovarian cancer	40 (0.6)	1 (0.8)	39 (0.6)	0.524
Endometrial carcinoma	22 (0.3)	2 (1.6)	20 (0.3)	0.061
Usage of GnRHa before surgery	1291 (19.0)	2 (1.6)	1289 (19.3)	**<0.001**
History gynecological surgery				**0.017**
Benign ovarian cystectomy	543 (8.0)	19 (15.2)	524 (7.8)	**0.003**

TABLE 1 (Continued)

Characteristic	Total cases ($n = 6809$)	EAOC ($n = 125$)	Non-EAOC ($n = 6684$)	p value
Oophorectomy	71 (1.0)	7 (5.6)	64 (1.0)	**<0.001**
Salpingectomy	121 (1.8)	6 (4.8)	115 (1.7)	**0.010**
Myomectomy	135 (2.0)	5 (4.0)	130 (1.9)	0.102
Adenomyomectomy	13 (0.2)	0 (0)	13 (0.2)	0.786
Transcervical resection of polyp	111 (1.6)	0 (0)	111 (1.7)	0.126
Hysterectomy	98 (1.4)	4 (3.2)	94 (1.4)	0.105
Tubal ligation	25 (0.4)	0 (0)	25 (0.4)	**0.017**
Serum CA125 (U/ml)	69.9 ± 151.4	241.9 ± 485.4	65.9 ± 131.6	**<0.001**
Serum CA199 (U/ml)	34.9 ± 115.6	173.9 ± 536.2	30.5 ± 64.6	**0.005**
Maximum diameter of tumor by US (cm)	5.2 ± 4.1	8.2 ± 4.5	5.2 ± 4.1	**<0.001**
Maximum diameter of tumor in surgery (cm)	5.1 ± 3.0	8.4 ± 5.4	5.1 ± 3.0	**0.015**
Tumor position in surgery				0.230
Left side	965 (14.2)	15 (12.0)	950 (14.2)	
Right side	520 (7.6)	6 (4.8)	526 (7.9)	
Both sides	325 (4.8)	3 (2.4)	322 (4.8)	
NA	4993 (73.3)	101 (80.8)	4892 (73.2)	
Nature of the tumor				**<0.001**
Cystic	570 (8.4)	10 (8.0)	560 (8.4)	
Solid	39 (0.6)	5 (4.0)	34 (0.5)	
Cystic-solid	183 (2.7)	30 (24.0)	153 (2.3)	
Thickness of endometrium (cm)	0.8 ± 0.5	0.7 ± 0.5	0.8 ± 0.5	0.079
Leiomyoma or adenomyoma showed in US	2122 (31.2)	29 (23.2)	2093 (31.3)	0.052
Gynecological complications				
Leiomyoma	1988 (29.2)	21 (16.8)	1967 (29.4)	**0.002**
Adenomyosis	1246 (18.3)	31 (24.8)	1215 (18.2)	0.058
Endometrial hyperplasia	29 (0.4)	1 (0.8)	28 (0.4)	0.416
EIN	55 (0.8)	3 (2.4)	52 (0.8)	0.080
Endometrial polyp	747 (11.0)	10 (8.0)	737 (11.0)	0.283
Ovarian borderline tumor	7 (0.1)	2 (1.6)	5 (0.1)	**0.007**
Subtypes of endometriosis				**<0.001**
Ovarian endometrioma	4313 (63.3)	63 (50.4)	4250 (63.6)	**0.002**
Abdominal endometriosis	537 (7.9)	26 (20.8)	511 (7.6)	**<0.001**
DIE	581 (8.5)	6 (4.8)	575 (8.6)	0.132
Distant endometriosis	398 (5.8)	1 (0.8)	397 (5.9)	**0.004**

Note: Data are presented as mean ± standard deviation, n (%), or median (range) unless otherwise indicated. The bold values are those less than 0.05.
Abbreviations: AUB, abnormal uterine bleeding; BMI, body mass index; CA125, carbohydrate antigen 125; CA199, carbohydrate antigen 199; CS, cesarean section; DIE, deep invasive endometriosis; EAOC, endometriosis-associated ovarian cancer; EIN, endometrial intraepithelial neoplasma; GnRHa, gonadotropin-releasing hormone agonist; NA, not available; SD, standard deviation; US, ultrasound; VD, vaginal delivery.

3.5 | Validation of the machine learning risk model

The same NLP method was used to extract the prospective dataset from September 1, 2019, to May 30, 2021, and 1788 of 2378 cases were finally enrolled for the validation. The machine learning risk model had an AUC of 0.8758, with a sensitivity of 94.4% and a specificity of 73.8% (Figure 6). To facilitate the practice used, we created a setup.exe to help predict the occurrence rate of EAOC in patients with endometriosis.

4 | DISCUSSION

The prevalence of EAOC was 1.84% in this study (125/6809), which was slightly higher than in the previous report.[2] In this study, we built a risk model through machine learning that could predict the risk of EAOC in patients with endometriosis, with a sensitivity of 86.8% and a specificity of 86.7%. The decision curve analysis showed that the machine learning-based risk model had greater benefit for the

TABLE 2 Variables enrolled in the logistics regression model

Variable	B	SE	Sig.	OR (95% CI)
Age at diagnosis	0.12	0.01	<0.001	1.130 (1.098–1.163)
Chief complaints of dysmenorrhea	−1.05	0.35	0.003	0.351 (0.176–0.700)
Usage of GnRHa before surgery	−1.95	0.74	0.009	0.143 (0.033–0.615)
Family history of EC	2.42	0.94	0.010	11.263 (1.770–71.662)
History of benign cystectomy	1.22	0.38	0.002	3.371 (1.589–7.154)
Mirena	1.90	0.99	0.054	6.717 (0.965–46.768)
Serum CA125 before surgery	0.001	0.001	<0.001	1. 002 (1.001–1.002)
Maximum diameter of tumor by US	0.03	0.01	0.042	1.027 (1.001–1.053)
Leiomyoma or adenomyoma indicated by US	−1.16	0.33	<0.001	0.312 (0.163–0.600)
Peritoneal endometriosis confirmed by pathology	2.23	0.37	<0.001	9.332 (4.546–19.157)
Constant	−8.62	0.73	<0.001	0.00

Abbreviations: CA125, carbohydrate antigen 125; CI, confidence interval; EC, endometrial carcinoma; GnRHa, gonadotropin-releasing hormone agonist; OR, odds ratio; US, ultrasound.

cohort of patients with endometriosis than the logistic regression risk model, and the difference was of statistical significance. During the validation process, the same NLP method was used to collect the prospective dataset, and the prevalence of EAOC was 1.01% (18/1788). The performance of the machine learning-based risk model was also validated using this prospective dataset, revealing a sensitivity of 94.4% and a specificity of 73.8%. To make this machine learning-based risk model more practical and convenient, we created a setup.exe that is easy to install and convenient for clinical use. The risk prediction of malignant transformation can be calculated with either a piece or a batch of data.

It has been proposed that factors predictive of endometriosis malignancy include increasing age, postmenopausal status, higher levels of carbohydrate antigen 125, larger endometriomas, and long-standing endometriosis.[16] A previous study in this center explored the risk factors for EAOC among 1038 women with endometrioma aged ≥45 years, revealing that risk factors included menopausal status, tumors ≥8 cm in diameter, and coexisting endometrial disorders.[17] Furthermore, infertility and nulliparity are reported to be associated with a higher risk of endometriosis-related ovarian cancer.[18–21] All in all, previous studies have described the difference of epidemiology and clinicopathologic characteristics between those with endometriosis and those with EAOC. However, no consensus was reached concerning these high-risk factors for EAOC in the previous studies, and no predictive risk model exists for clinical practice.

To distinguish malignant transformation of endometriosis, increasing differential methods have been proposed or explored, ranging from the most common and cost-effective imaging examinations to innovative approaches, including biochemical and technical advancement. The use of magnetic resonance relaxometry as a noninvasive tool to help discriminate EAOC from ovarian endometrioma had a sensitivity of 86% and specificity of 94%.[22,23] The sensitivity of serum Smac + HE4 + CA125 was the highest, at up to 98.3%, although specificity was only up to 75%.[24] A relation between methylation status of RASSF2A gene promoter and EAOC has also been revealed. The combination of miR-16, 21, and 191 may represent a signature unique to EAOCs.[25] Furthermore, proteomic analysis of endometrial fluid and circulating tumor DNA may also be used to detect precursor lesions in EAOC and to investigate the risk of developing EAOC.[26] The total iron levels of cyst fluid may have helped discriminate EAOC from ovarian endometrioma with a sensitivity of 85% and a specificity of 98% when the cut-off point of total iron was set at 64.8 mg/L.[22] However, the application of these study results was limited by their small sample size, the difficulty of obtaining samples to test, high costs, simultaneous confounding factors, shortage of satisfactory accuracy, or poor clinical practicality. More importantly, the results of previous studies lacked prospective validation. Thus, more practical predictive methods still need to be explored.

The early and accurate prediction of malignant transformation of endometriosis remains challenging. More research is urgently needed to establish a reliable risk model for predicting EAOC in patients with endometriosis.[27] To our delight, machine learning has been widely used in clinical differentiation and early diagnosis, such as Parkinson's disease,[28] diabetic kidney disease,[29] non-alcoholic fatty liver disease,[30] and prostate cancer diagnosis.[31] Compared with traditional methods, the use of machine learning algorithms has advantages for modeling and validation. A random subsampling scheme was used to minimize the estimated bias. Different clinicopathological characteristics have been described in the previous studies between patients with EAOC and those with endometriosis. However, the clinical data were not fully used to predict the occurrence of endometriosis malignancy. In this study, we applied machine learning to construct a risk model to predict endometriosis malignancy, which was the novelty of this pilot study. This model includes not only clinical risk factors related to EAOC, which have been previously reported, but also unreported clinicopathologic factors as it was constructed making full use of these clinical variables.

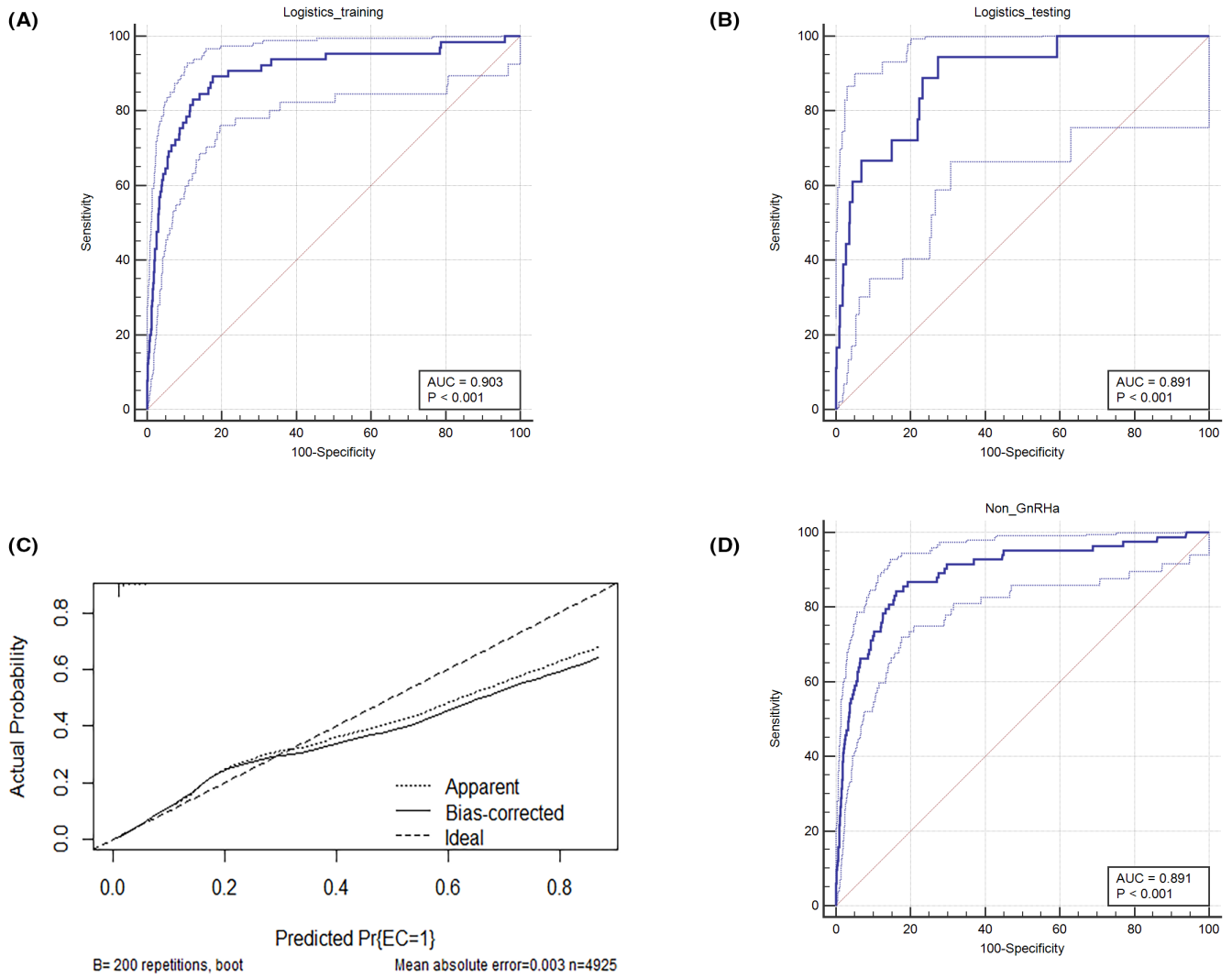

FIGURE 2 Receiver operating characteristic curve of the logistic regression model for the training dataset (A) and testing dataset (B). (C) Calibration curve of the logistic regression model. (D) Receiver operating characteristic curve of the logistic regression model for patients not using gonadotrophin-releasing hormone agonists (GnRHa). AUC, area under the receiver operating characteristic curve.

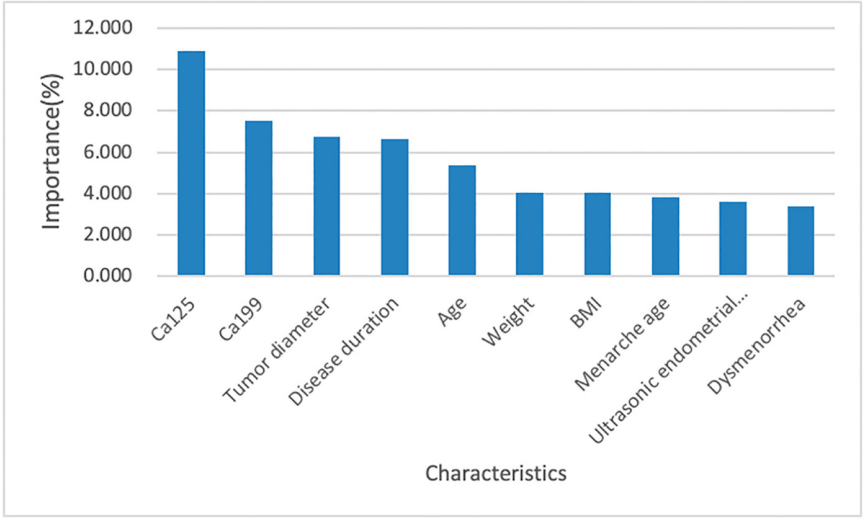

FIGURE 3 The features of importance for each variable enrolled in the logistic regression model. BMI, body mass index.

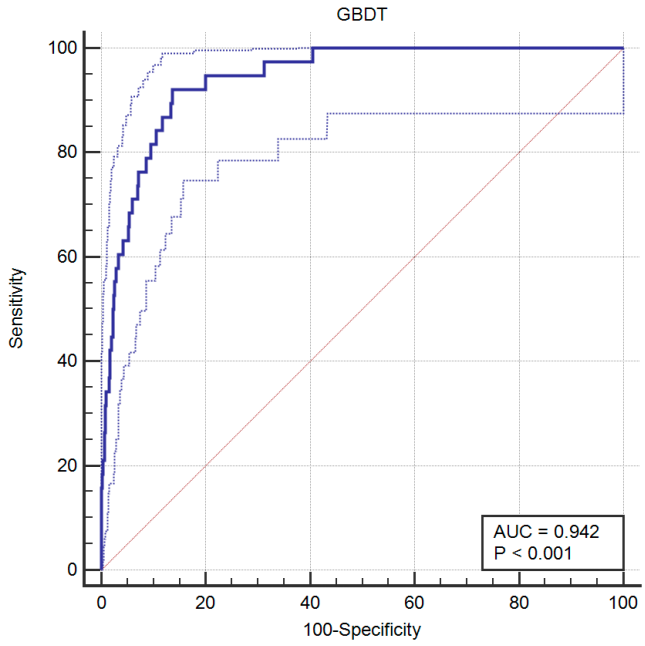

FIGURE 4 Receiver operating characteristic curve of the gradient-boosting machine with the most optimal hyperparameters. GBDT, gradient-boosting decision tree.

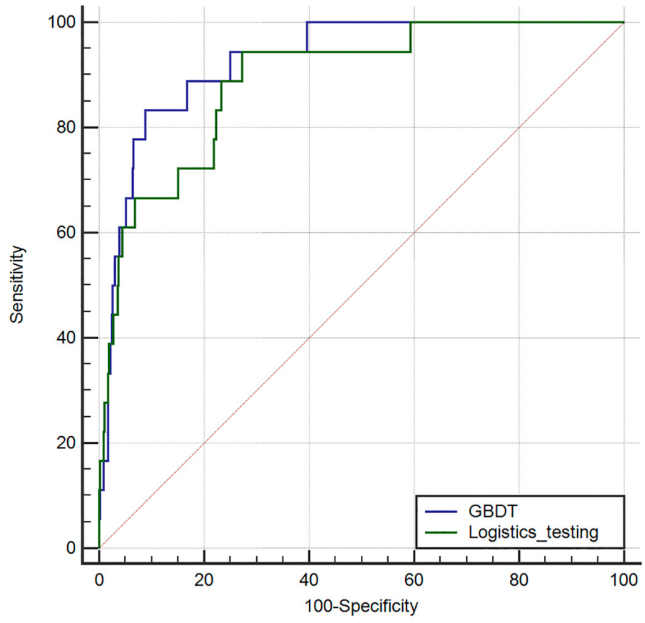

FIGURE 5 The receiver operating characteristic curve comparison of the two-risk model for endometriosis-associated ovarian cancer. GBDT, gradient-boosting decision tree.

The results from this pilot study showed that the machine learning-based risk model has a satisfactory predictive effect, and it should be noted that a setup.exe was constructed to facilitate its clinical application.

This study has both strengths and limitations. The machine learning-based risk model is of great generalizability for its features. Although this is only a single tertiary center analysis, the sample size

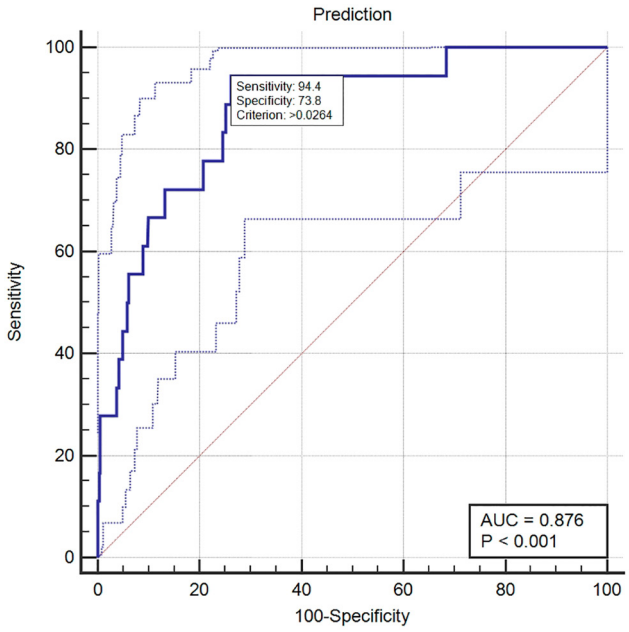

FIGURE 6 Receiver operating characteristic curve of the gradient boosting machine with the most optimal hyperparameters. AUC, area under the receiver operating characteristic curve.

is large enough and was validated in a prospective dataset. The risk model can provide an occurrence rate for EAOC, which previous research could not. The construction of a setup.exe is also a highlight, as this facilitates clinical practice and may even become a reference tool for primary medical institutions. However, its limitations should not be ignored. Although repeated cross-validations were used to minimize bias, over-fitting and higher accuracy results may remain. The lack of external validation with independent datasets from other centers for generalization and the biases inherent in retrospective studies and cross-sectional analysis are additional limitations.

5 | CONCLUSION

A risk model based on machine learning was constructed to help predict which patients were most likely to experience malignant transformation of endometriosis. It demonstrates that clinical characteristics can be applied to facilitate the early diagnosis of EAOC. This risk model can help gynecologists recognize patients at potential risk of EAOC and help with the monitoring and pursuing of risk-reducing medical or surgical treatment regimens. However, the model still requires further validation with a larger, prospective, multicenter dataset to provide favorable evidence for its clinical application.

AUTHOR CONTRIBUTIONS

SW and JL: Conception and design. SW and QF: Provision of study materials or patients. CX: Collection and assembly of data. XC and SW: Data analysis and interpretation. All authors: Manuscript writing and final manuscript approval.

FUNDING INFORMATION

This study was funded by the National Key R&D Program of China (No. 2017YFC1001200) and the CAMS Innovation Fund for Medical Sciences (CIFMS) (No. CIFMS 2021-I2M-1-003).

CONFLICT OF INTEREST

None.

ORCID

Shu Wang https://orcid.org/0000-0001-5447-0946

REFERENCES

1. Shafrir AL, Farland LV, Shah DK, et al. Risk for and consequences of endometriosis: a critical epidemiologic review. *Best Pract Res Clin Obstet Gynaecol*. 2018;51:1-15.
2. Bulun SE, Yilmaz BD, Sison C, et al. Endometriosis. *Endocr Rev*. 2019;40:1048-1079.
3. McCluggage WG. Endometriosis-related pathology: a discussion of selected uncommon benign, premalignant and malignant lesions. *Histopathology*. 2020;76:76-92.
4. Bogani G, Ditto A, Lopez S, et al. Adjuvant chemotherapy vs. observation in stage I clear cell ovarian carcinoma: a systematic review and meta-analysis. *Gynecol Oncol*. 2020;157:293-298.
5. Bogani G, Tagliabue E, Ditto A, et al. Assessing the risk of pelvic and Para-aortic nodal involvement in apparent early-stage ovarian cancer: a predictors- and nomogram-based analyses. *Gynecol Oncol*. 2017;147:61-65.
6. Paik ES, Kim TJ, Choi CH, Kim BG, Bae DS, Lee JW. Clinical outcomes of patients with clear cell and endometrioid ovarian cancer arising from endometriosis. *J Gynecol Oncol*. 2018;29:e18.
7. Wang S, Qiu L, Lang JH, et al. Clinical analysis of ovarian epithelial carcinoma with coexisting pelvic endometriosis. *Am J Obstet Gynecol*. 2013;208(413):e411-e415.
8. Force USPST, Grossman DC, Curry SJ, et al. Screening for ovarian cancer: US preventive services task Force recommendation Statement. *Jama*. 2018;319:588-594.
9. Murakami K, Kotani Y, Nakai H, Matsumura N. Endometriosis-associated ovarian cancer: the origin and targeted therapy. *Cancers (Basel)*. 2020;12:1676.
10. Kumar S, Munkarah A, Arabi H, et al. Prognostic analysis of ovarian cancer associated with endometriosis. *Am J Obstet Gynecol*. 2011;204(63):e61-e67.
11. Scarfone G, Bergamini A, Noli S, et al. Characteristics of clear cell ovarian cancer arising from endometriosis: a two center cohort study. *Gynecol Oncol*. 2014;133:480-484.
12. Van Gorp T, Amant F, Neven P, Vergote I, Moerman P. Endometriosis and the development of malignant tumours of the pelvis. A review of literature. *Best Pract Res Clin Obstet Gynaecol*. 2004;18:349-371.
13. Voytovich L, Greenberg C. Natural language processing: practical applications in medicine and investigation of contextual autocomplete. *Acta Neurochir Suppl*. 2022;134:207-214.
14. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*. 1988;44:837-845.
15. Coutant C, Olivier C, Lambaudie E, et al. Comparison of models to predict nonsentinel lymph node status in breast cancer patients with metastatic sentinel lymph nodes: a prospective multicenter study. *J Clin Oncol*. 2009;27:2800-2808.
16. Dahiya A, Sebastian A, Thomas A, George R, Thomas V, Peedicayil A. Endometriosis and malignancy: the intriguing relationship. *Int J Gynaecol Obstet*. 2021;155:72-78.
17. He ZX, Shi HH, Fan QB, et al. Predictive factors of ovarian carcinoma for women with ovarian endometrioma aged 45 years and older in China. *J Ovarian Res*. 2017;10:45.
18. Ness RB. Endometriosis and ovarian cancer: thoughts on shared pathophysiology. *Am J Obstet Gynecol*. 2003;189:280-294.
19. Brinton LA, Lamb EJ, Moghissi KS, et al. Ovarian cancer risk associated with varying causes of infertility. *Fertil Steril*. 2004;82:405-414.
20. Melin A, Sparén P, Bergqvist A. The risk of cancer and the role of parity among women with endometriosis. *Hum Reprod*. 2007;22:3021-3026.
21. Brinton LA, Westhoff CL, Scoccia B, et al. Causes of infertility as predictors of subsequent cancer risk. *Epidemiology*. 2005;16:500-507.
22. Yoshimoto C, Takahama J, Iwabuchi T, Uchikoshi M, Shigetomi H, Kobayashi H. Transverse relaxation rate of cyst fluid can predict malignant transformation of ovarian endometriosis. *Magn Reson Med Sci*. 2017;16:137-145.
23. Kobayashi H, Yamada Y, Kawahara N, Ogawa K, Yoshimoto C. Modern approaches to noninvasive diagnosis of malignant transformation of endometriosis. *Oncol Lett*. 2019;17:1196-1202.
24. Xu XR, Wang X, Zhang H, Liu MY, Chen Q. The clinical significance of the combined detection of serum Smac, HE4 and CA125 in endometriosis-associated ovarian cancer. *Cancer Biomark*. 2018;21:471-477.
25. Suryawanshi S, Vlad AM, Lin HM, et al. Plasma microRNAs as novel biomarkers for endometriosis and endometriosis-associated ovarian cancer. *Clin Cancer Res*. 2013;19:1213-1224.
26. Dawson A, Fernandez ML, Anglesio M, Yong PJ, Carey MS. Endometriosis and endometriosis-associated cancers: new insights into the molecular mechanisms of ovarian cancer development. *Ecancermedicalscience*. 2018;12:803.
27. Hermens M, van Altena AM, Nieboer TE, et al. Incidence of endometrioid and clear-cell ovarian cancer in histological proven endometriosis: the ENOCA population-based cohort study. *Am J Obstet Gynecol*. 2020;223:107.e101-107.e111.
28. Zhang J. Mining imaging and clinical data with machine learning approaches for the diagnosis and early detection of Parkinson's disease. *NPJ Parkinsons Dis*. 2022;8:13.
29. Allen A, Iqbal Z, Green-Saxena A, et al. Prediction of diabetic kidney disease with machine learning algorithms, upon the initial diagnosis of type 2 diabetes mellitus. *BMJ Open Diabetes Res Care*. 2022;10:e002560.
30. Ma X, Yang C, Liang K, et al. A predictive model for the diagnosis of non-alcoholic fatty liver disease based on an integrated machine learning method. *Am J Transl Res*. 2021;13:12704-12713.
31. Chiu PK, Shen X, Wang G, et al. Enhancement of prostate cancer diagnosis by machine learning techniques: an algorithm development and validation study. *Prostate Cancer Prostatic Dis*. 2021:1-5.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Chao X, Wang S, Lang J, Leng J, Fan Q. The application of risk models based on machine learning to predict endometriosis-associated ovarian cancer in patients with endometriosis. *Acta Obstet Gynecol Scand*. 2022;101:1440-1449. doi: 10.1111/aogs.14462