Am J Cancer Res 2024;14(6):3186-3197
www.ajcr.us /ISSN:2156-6976/ajcr0156567

Original Article

Developing a comprehensive molecular subgrouping model for cervical cancer using machine learning

Gwan Hee Han[1], Hae-Rim Kim[2], Hee Yun[3], Joon-Yong Chung[4], Jae-Hoon Kim[5,6], Hanbyoul Cho[5,6]

[1]Department of Obstetrics and Gynecology, Sanggye Paik Hospital, Inje University College of Medicine, Seoul 01757, Republic of Korea; [2]Department of Statistics, College of Natural Science, University of Seoul, Seoul 02504, Republic of Korea; [3]Department of Obstetrics and Gynecology, Gangnam Severance Hospital, Yonsei University College of Medicine, Seoul 06299, Republic of Korea; [4]Molecular Imaging Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA; [5]Department of Obstetrics and Gynecology, Yonsei University College of Medicine, Seoul 03722, Republic of Korea; [6]Institute of Women's Life Medical Science, Yonsei University College of Medicine, Seoul 03722, Republic of Korea

Received March 14, 2024; Accepted June 17, 2024; Epub June 25, 2024; Published June 30, 2024

Abstract: This study developed a molecular classification model for cervical cancer using machine learning, integrating prognosis related biomarkers with clinical features. Analyzing 281 specimens, 27 biomarkers were identified, associated with recurrence and treatment response. The model identified four molecular subgroups: group 1 (OALO) with Overexpression of ATP5H and LOw risk; group 2 (LASIM) with low expression of ATP5H and SCP, indicating InterMediate risk; group 3 (LASNIM) characterized by Low expression of ATP5H, SCP, and NANOG, also at InterMediate risk; and group 4 (LASONH), with Low expression of ATP5H, and SCP, Over expression of NANOG, indicating High risk, and potentially aggressive disease. This classification correlated with clinical outcomes such as tumor stage, lymph node metastasis, and response to treatment, demonstrating that combining molecular and clinical factors could significantly enhance the prediction of recurrence and aid in personalized treatment strategies for cervical cancer.

Keywords: Artificial intelligence, cervical cancer, machine learning, prognosis

Introduction

Cervical cancer is a significant global health concern, ranking fourth in both incidence and mortality rate worldwide, with over 600,000 new cases and 341,843 death in 2022 [1]. In Korea, despite a decrease in incidence due to screening and human papillomavirus (HPV) vaccination, it remains a leading cause of cancer-related deaths [2]. While early stage cervical cancer can be effectively treated with radical hysterectomy with adjuvant concurrent chemoradiation therapy (CCRT), recurrence rate are high, posing a challenge to long-term survival. In this context, precise patient selection and prognostic prediction are paramount for enhancing clinical outcomes.

Traditionally, the International Federation of Gynecology and Obstetrics (FIGO) staging system has been the cornerstone for evaluating prognosis and guiding therapeutic decisions in cervical cancer management. However, the advent of precision medicine has highlighted the limitations of the FIGO staging system in accounting for the individual variabilities in patient outcomes. Although, the FIGO 2018 classification made significant strides by incorporating pathologic and imaging evidence, it still does not capture the full spectrum of prognostic factors [3]. Tissue biomarkers, for instance, offer invaluable insights into the pathogenic processes and response to therapeutic interventions but are not adequately represented in the current staging system [4]. Furthermore, the molecular landscape of cervical cancer, characterized by its heterogeneity, plays a pivotal role in influencing treatment responses and disease progression. The activation and deactivation of proto-oncogenes and tumor suppressor genes, often though the integration of high-risk HPV DNA into the host genome, are

https://doi.org/10.62347/MTER1763

critical in the pathogenesis of cervical cancer. This complexity underscores the need for prognostic models that can harness the wealth of clinical and molecular data to predict patient outcomes more accurately [5].

While significant process has been made in other types of cancer, cervical cancer lacks comprehensive prognostic models that amalgamate clinical factors with pathological and molecular insights [6-8]. The heterogeneity within cervical cancer poses unique challenges in treatment and prognosis, necessitating a multifaced approach to understanding and managing the disease. Studies such as the one conducted by Muñoz *et al.,* have begun to shed light on the intricacies of cervical cancer heterogeneity, advocating for more integrated approach to prognostic modeling [9].

This backdrop sets the stage for our study, which aims to bridge the gap in cervical cancer prognosis by developing a molecular classification system. By leveraging advanced machine learning algorithms and integrating a diverse array of biomarkers, our goal is to provide a more nuanced and predictive framework for patient stratification, ultimately guiding personalized treatment plans and improving patient survival rate in the face of cervical cancer's complexity.

Material and methods

Patients and immunohistochemistry

The study analyzed 205 cervical cancer tissues collected from patients at Gangnam Severance Hospital from 1996 to 2010, used for tissue microarray (TMA) construction and subsequent immunohistochemistry (IHC) analysis. These tissues were selected after thorough review by a gynecological pathologist. Patient data including demographics, tumor characteristics, treatment outcomes, and survival data were compelled. The study adhered to FIGO2018 and WHO classification systems for staging and grading, with specific exclusion criteria to maintain data integrity [10, 11]. We utilized the previously collected data for IHC of ATP5H, SCP3, NANOG, FOXO1, PAX3, HSP90a, pEGFR, CRY1, HIF-1α, TRPV1, and pAKT, in the cytoplasm and pERK1/2 in the nucleus for further investigation of the molecular classification [12-21]. All procedures were conducted in accordance with the Declaration of Helsinki, and the study was approved by the Institutional Review Board of Gangnam Severance Hospital (approval no. #3-2021-0496).

Development of molecular classification for cervical cancer

We segmented the dataset into training (70%) and test (30%) sets, further dividing the training set into derivation and validation subsets. Out model construction utilized 27 protein expression levels, applying a Cox proportional hazard model with a LASSO penalty to establish a novel molecular classification. This methods was refined to optimized the penalization parameter (λ) to enhance the model's predictive accuracy. A comparison between linear and non-linear models (the latter developed using XGBoost) revealed comparable performance, with further details and statistical analysis presented in the Supplementary Figure 1.

Statistical analysis

Baseline characteristics were presented using descriptive statistics, with continuous variable compared using t-test and Mann-Whitney U test, and categorical variables using Chi-squared and Fisher's exact tests. Survival analysis was conducted using univariate and multivariate Cox proportional hazards models, with the prognostic accuracy of new and conventional predictors compared using Harrell's C-index. Progression-free survival (PFS) was analyzed using the Kaplan-Meier method, with the log-rank test assessing the significance of differences in survival predicted by new classification systems. Statistical significance was determined at *P*-values less than 0.05. All statistical analyses were performed using R (version 4.3.1).

Results

Patient characteristics

We present the baseline characteristics of 271 cervical cancer patients, divided into training (70%) and test (30%) sets to develop a prognostic model. The test set included patients with stage I-IIA (64 patients) and IIB-IV (25 patients) cancer, with a median age of 50.55 years. Squamous cell carcinoma was predominant (75.28%), followed by other types (24.72%).

Table 1. Baseline characteristics of cervical cancer patients

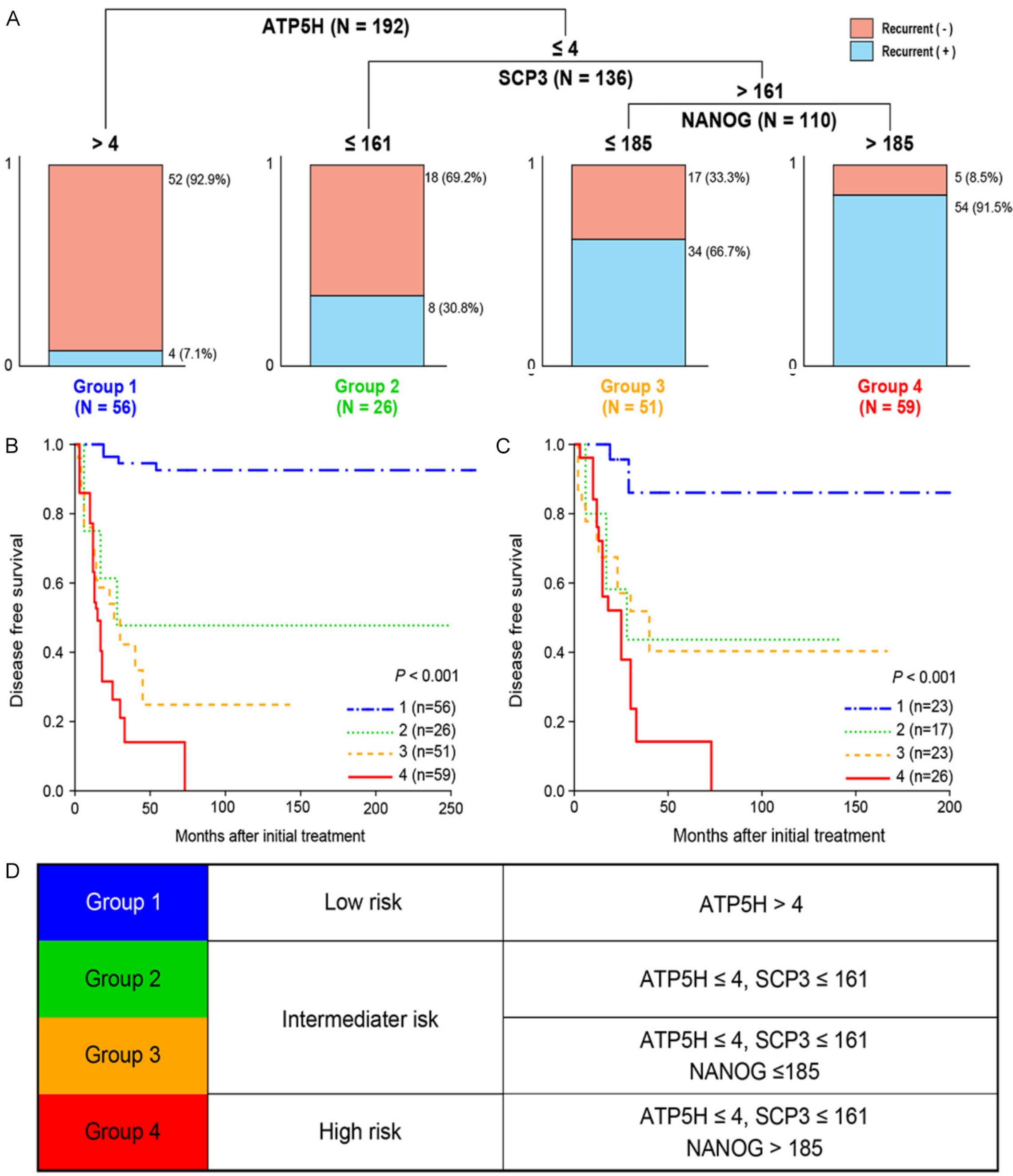
	All (N = 281)	Test cohort (N = 89)	Training cohort (N = 192)	*P*-value
Age, years old	49.37 ± 11.40	50.55 ± 10.76	48.82 ± 11.66	0.236
FIGO stage				0.363
I-IIA	190 (67.62%)	64 (71.91%)	126 (65.62%)	
IIB-IV	91 (32.38%)	25 (28.09%)	66 (34.38%)	
Cell Type				0.097
Squamous	229 (81.49%)	67 (75.28%)	162 (84.38%)	
Others	52 (18.51%)	22 (24.72%)	30 (15.62%)	
Grade				0.189
Low/Moderate	158 (60.77%)	57 (67.06%)	101 (57.71%)	
Poor	102 (39.23%)	28 (32.94%)	74 (42.29%)	
Tumor size (mm)	30.54 ± 19.40	28.54 ± 17.04	31.47 ± 20.38	0.24
LVI				0.934
Negative	95 (51.35%)	28 (50.00%)	67 (51.94%)	
Positive	90 (48.65%)	28 (50.00%)	62 (48.06%)	
Lymph node metastasis				0.911
Negative	168 (73.68%)	50 (72.46%)	118 (74.21%)	
Positive	60 (26.32%)	19 (27.54%)	41 (25.79%)	
Response to Chemoradiation therapy				0.293
Good	81 (57.45%)	31 (64.58%)	50 (53.76%)	
Poor	60 (42.55%)	17 (35.42%)	43 (46.24%)	

FIGO, International Federation of Gynecology and Obstetrics; LN metastasis, Lymph node metastasis.

Most cases were low/moderate grade (60.77%), with a minority being poor grade (32.94%). The median tumor size was 28.54 mm, and lymph node (LN) metastasis was observed in 27.54% of patients. No significant differences were found between the training and test sets in baseline characteristics, including response to CCRT (**Table 1**).

Stratified prognostic biomarker assessment in cervical cancer prognosis

Our investigation further progressed into a sophisticated molecular dimension where we employed lasso-penalized Cox regression within the derivation subset of the training cohort. The choice of lasso penalization was pivotal, as it simultaneously facilitated variable selection and regularization, helping to enhance the predictive accuracy of model while preventing overfitting. By focusing on minimizing the partial likelihood deviation, we identified a core set of seven biomarkers through this reinforced statistical approach. These biomarkers underwent a rigorous validation process via bootstrap resampling, ensuring their robustness in the model. Notable ATP5H was a consistent feature in all models, underscoring its potential as a key predictive biomarker, while SCP3, NANOG, pERK1/2 and FOXO1 were recurrent in over 95% of the models, highlighting their significant prognostic relevance. Following this, we adjusted for age difference in our cohort, which led to the selection of ATP5H, SCP3, and NANOG as the primary biomarkers for our final molecular classification (Supplementary Figure 2; Supplementary Table 1) [22]. The decision tree was fined tuned using entropy-based measure, aiming to maximize the information gain at each decision node. This strategic application yields specific cut-off points for patient stratifications, effectively segregating patients into distinct prognostic categories. As a result, the cut off value was determined, and grouping was performed based on the trained model. As shown in **Figure 1A**, the molecular classification of the 192 fully evaluated cases yielded four different subgroups. (1) The first group consisted of patients who overexpressed ATP5H (ATP5H > 4, which is the cutoff value). (2) The second group included patients who expressed both ATP5H and SCP3 at a level lower than the cutoff value (ATP5H ≤ 4; SCP ≤ 161). The third group comprised patients who



Group 1	Low risk	ATP5H > 4
Group 2	Intermediater isk	ATP5H ≤ 4, SCP3 ≤ 161
Group 3		ATP5H ≤ 4, SCP3 ≤ 161 NANOG ≤185
Group 4	High risk	ATP5H ≤ 4, SCP3 ≤ 161 NANOG > 185

Figure 1. Molecular classification for predicting cervical cancer recurrence. (A) Decision tree with cut off point for molecular classification was visualized graphically. Kaplan-Meier survival analysis according to molecular classification by disease free survival (DFS) of (B) training cohort and (C) test cohort. (D) A molecular classification stratified as low, intermediate, and high-risk group.

expressed ATP5H, SCP3, and NANOG at levels lower than the cutoff value (ATP5H ≤ 4; SCP ≤ 161; NANOG ≤ 185). Finally, the fourth group comprised patients who expressed ATP5H and SCP3 at levels lower than the cutoff value but overexpressed NANOG than the cutoff value (ATP5H ≤ 4; SCP ≤ 161; NANOG > 185; **Figure 1A**, Supplementary Figure 3). Subsequently, Kaplan-Meier analysis was performed to optimize risk stratification and visualize the differences in PFS between the identified risk subgroups. The analysis revealed that the low-risk group (Group 1: OALO - Overexpression of ATP5H and Low risk) exhibited the most favor-

Table 2A. Cox regression of training dataset

Group	N	Event (%)	Hazard ratio (95% CI)	*P*-value
1	56	4 (7.14%)	Ref	
2	26	8 (30.77%)	11.63 (3.5-38.68)	< 0.001
3	51	34 (66.67%)	19.73 (6.96-55.93)	< 0.001
4	59	54 (91.53%)	34.4 (12.3-96.21)	< 0.001

Table 2B. Cox regression of test dataset

Group	N	Event (%)	Hazard ratio (95% CI)	*P*-value
1	23	3 (13.04%)	Ref	
2	17	8 (47.06%)	6.71 (1.78-25.38)	< 0.001
3	23	12 (52.17%)	6.69 (1.89-23.72)	< 0.001
4	26	23 (88.46%)	14.29 (4.24-48.16)	< 0.001

able outcome, while groups 2 and 3 (LASIM - Low expression of ATP5H and SCP and Intermediate risk; LASNIM - Low expression of ATP5H, SCP, and NANOG and Intermediate risk, respectively) were classified as intermediate-risk, and group 4 (LASONH - Low expression of ATP5H and SCP, Overexpression of NANOG, and High risk) was deemed high-risk. The PFS values demonstrated a significant difference among the three risk subgroups in both the training and test cohorts, indicating the effectiveness of the molecular classification model in accurately stratifying patients with cervical cancer according to their risk level (Both *P* < 0.001; **Figure 1B-D**). After identifying four distinct subgroups of cervical cancer based on the gene expression data, we evaluated their performance using Cox regression analysis during the training of the test cohorts. Our results were statistically significant and revealed a clear trend among the subgroups. Specifically, group 1 demonstrated the most favorable outcome with the lowest HR, whereas the HR values increased progressively from groups 1 to 4, indicating a worsening prognosis from groups 1 to 4. Notably, group 4 exhibited the highest HR and the poorest outcome among all the subgroups (HR = ref; HR = 11.63, 95% CI = 3.5-38.68, *P* < 0.001; HR = 19.73, 95% CI = 6.96-55.9, *P* < 0.001; HR = 34.4, 95% CI = 12.3-96.21, *P* < 0.001; **Table 2A**). Moreover, the test cohort showed the same result as the training cohort (**Table 2B**). These findings underscore the importance of identifying and characterizing distinct subgroups of cervical cancer using biomarker - biomarker interactions, and the addition of new prognostic biomarkers could improve cancer prognosis, as it may ultimately result in the development of personalized treatment strategies and improve patient outcomes.

Clinicopathological characteristics of the newly developed classification model

To comprehensively assess the benefits of the molecular subgroups in cervical cancer, we analyzed the clinicopathological characteristics of each subgroup (**Table 3**). Our results indicated that the squamous cell subtype was significantly more prevalent than the other subtypes across all subgroups (*P* = 0.045). Interestingly, we observed a significant advancement in FIGO staging from stage IIB to stage IV in groups 1 to 4, comprising 7.14%, 15.38%, 50.98%, and 52.24% of the cases, respectively. Additionally, tumor size significantly increased from 21.93 ± 20.31 mm to 40.25 ± 17.21 mm in patients from groups 1 to 4, respectively. Furthermore, lymph node metastasis and poor response to chemoradiotherapy were significantly more frequent in group 4 (61.54% and 59.62%, respectively) than in groups 1, 2, and 3. The prevalence of LVSI was significantly higher in groups 3 and 4 at 80% and 60%, respectively, than in other groups. Additionally, we compared the data for groups 2 and 3, as these groups were categorized as the intermediate group (**Table 4**). We observed that patients in group 3 were younger when diagnosed, were diagnosed at more advanced FIGO stages, had larger tumor sizes, and had increased LVSI compared to those in patients in group 2 (*P* = 0.027, *P* < 0.001, *P* < 0.001, *P* = 0.044, and *P* = 0.018, respectively; **Table 4**). Although these findings were statistically insignificant, a high incidence of lymph node metastasis was observed in group 3 (*P* = 0.050; **Table 4**).

Overall, the identification of distinct clinicopathological characteristics in each molecular subgroup of cervical cancer based on tumor size, lymph node metastasis, FIGO stage, LVSI, and response to chemoradiotherapy is important in cases where decisions need to be made regarding administration of further adjuvant

Table 3. Clinicopathological characteristics according to the new classification

	Group 1 (N = 56)	Group 2 (N = 26)	Group 3 (N = 51)	Group 4 (N = 59)	*P*-value
Age	47.55 ± 12.51	52.92 ± 11.29	47.18 ± 10.15	49.63 ± 12.00	0.159
FIGO stage					< 0.001
I-IIA	52 (92.86%)	22 (84.62%)	25 (49.01%)	27 (45.76%)	
IIB-IV	4 (7.14%)	4 (15.38%)	26 (50.98%)	32 (54.24%)	
Cell Type					0.045
Squamous	47 (83.9%)	19 (73.1%)	44 (86.3%)	52 (88.1%)	
Others	9 (16.1%)	7 (26.9%)	7 (13.7%)	7 (11.9%)	
Grade					0.462
Low/Moderate	31 (64.58%)	14 (63.64%)	29 (56.86%)	27 (50.00%)	
Poor	17 (35.42%)	8 (36.36%)	22 (43.14%)	27 (50.00%)	
Tumor size (mm)	21.93 ± 20.31	22.42 ± 16.62	36.92 ± 19.48	40.25 ± 17.21	< 0.001
LVSI					0.003
Negative	33 (68.8%)	15 (57.7%)	3 (20.0%)	16 (40.0%)	
Positive	15 (31.2%)	11 (42.3%)	12 (80.0%)	24 (60.0%)	
Lymph node metastasis					
Negative	45 (86.54%)	23 (88.46%)	18 (62.07%)	20 (38.46%)	0.004
Positive	7 (13.46%)	3 (11.54%)	11 (37.93%)	32 (61.54%)	
Response to Chemoradiation therapy					
Good	13 (92.86%)	8 (100.00%)	11 (57.89%)	21 (40.38%)	< 0.001
Poor	1 (7.14%)	0 (0.0%)	8 (42.11%)	31 (59.62%)	

FIGO, International Federation of Gynecology and Obstetrics; LVSI, Lymph vascular space invasion. Protein expression was determined through analysis of an immunohistochemically stained tissue array as described in the materials and methods section.

therapy and for predicting the prognosis of these patients. Therefore, our findings provide important insights that can guide the development of tailored treatment plans for patients and improve prediction of patient outcomes, ultimately improving the prognosis of patients with cervical cancer.

Advancement in prognostic modeling through clinic-molecular integration

To investigate whether the new molecular classification model could improve the predictive power for prognosis of cervical cancer, we developed an extended model by incorporating clinical information, such as FIGO stage, tumor size, LN metastasis, tumor grade, and age, into our new molecular classification model and compared the C-indices of the models containing clinical information alone and those containing the clinico-molecular classification (**Figure 2**). Remarkably, our results demonstrated that the combined clinico-molecular model, which considers both the new molecular classification and the clinical factors, could more accurately predict the prognosis than the model containing clinical information or the new molecular classification alone in test cohorts (mean C-indices 0.728, ranges 0.653-0.803, and $P < 0.001$; **Figure 2A**). Further elucidation is provided by **Figure 2B**, which portrays the model's performance overtime, measured by the AUC. The clinico-molecular model consistently outstripped the clinical only models showcasing its superior predictive consistency, particularly over an extended follow up period. This integrative model synthesizing molecular data with established clinical factors, emerges as a significantly more accurate prognostic tool, as evidenced by our statistical analyses. It underscores the importance of a dual faceted approach to risk stratification in cervical cancer paving the way for individualized treatment regimens and enhancing the granularity of clinical decision making.

Discussion

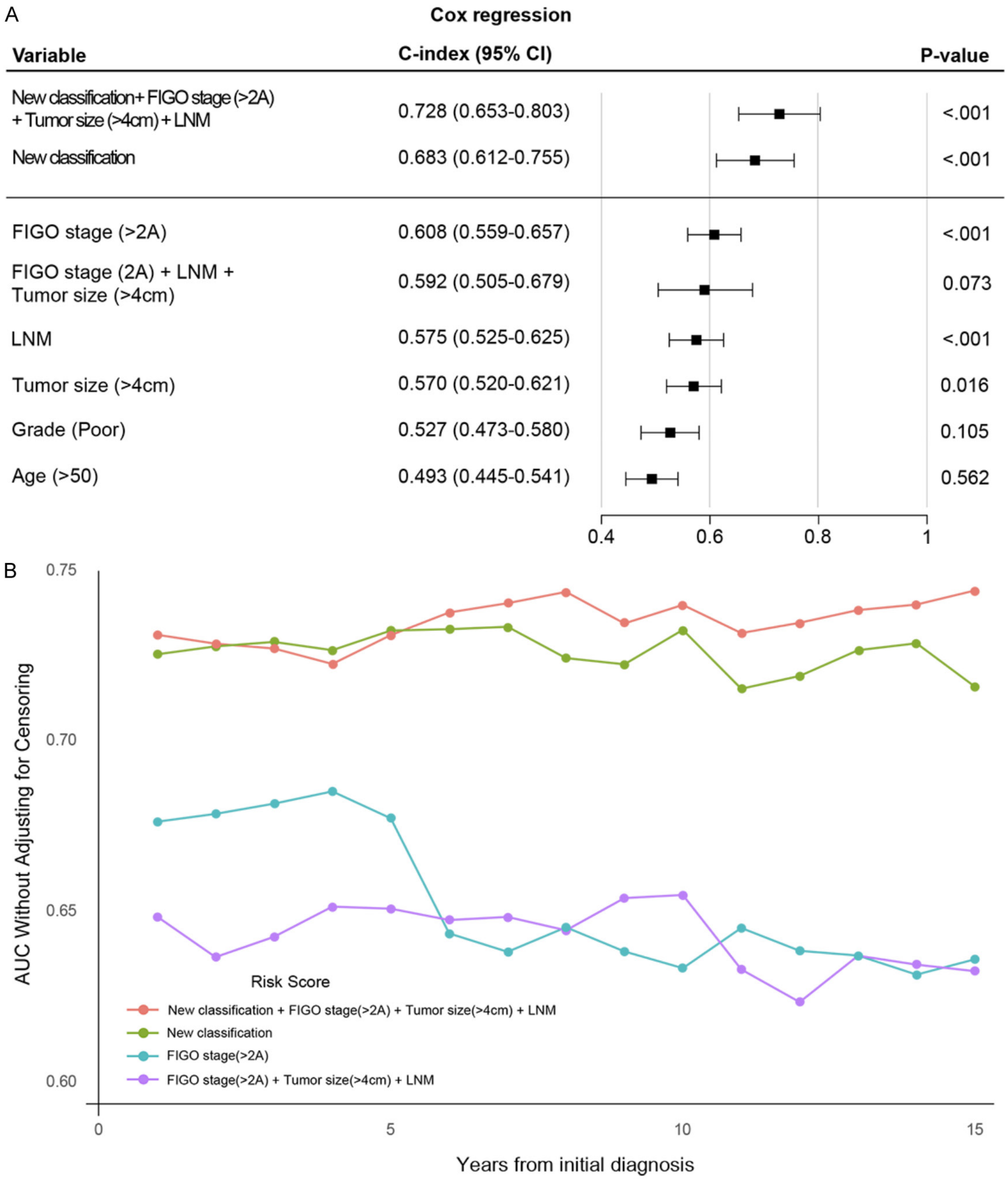
Prognosis prediction plays a critical role in clinical oncology as it provides valuable guidance

Table 4. Clinicopathological characteristics comparing group 2 and group 3

	Group 2 (N = 26)	Group 3 (N = 51)	*P*-value
Age	52.92 ± 11.29	47.18 ± 10.15	0.027
FIGO stage			< 0.001
I-IIA	22 (84.62%)	25 (49.01%)	
IIB-IV	4 (15.38%)	26 (50.98%)	
Cell Type			0.099
Squamous	19 (73.1%)	44 (86.3%)	
Others	7 (26.9%)	7 (13.7%)	
Grade			0.779
Low/Moderate	14 (63.64%)	29 (56.86%)	
Poor	8 (36.36%)	22 (43.14%)	
Tumor size (mm)	22.42 ± 16.62	36.92 ± 19.48	< 0.001
LVSI			0.044
Negative	15 (57.7%)	3 (20.0%)	
Positive	11 (42.3%)	12 (80.0%)	
LN metastasis			0.050
Negative	23 (88.46%)	18 (62.07%)	
Positive	3 (11.54%)	11 (37.93%)	
Response to chemoradiation therapy			0.018
Good	8 (100.00%)	11 (57.89%)	
Poor	0 (0.0%)	8 (42.11%)	

FIGO, International Federation of Gynecology and Obstetrics; LVSI, Lymph vascular space invasion; LN, Lymph node. Protein expression was determined through analysis of an immunohistochemically stained tissue array as described in the materials and methods section.

for treatment decisions, disease progression assessment, and patient management. However, while several studies have successfully applied artificial intelligence (AI) technology to develop and validate a molecular classification system by unveiling hidden and embedded patterns to predict prognosis and to provide valuable information regarding the response to therapy in various cancer types, including breast cancer [23], colorectal cancer [24], and sarcoma [25], limited research has specifically focused on cervical cancer. Therefore, in the present study, we used AI technology, specifically an ML algorithm, to develop and validate a molecular classification system for cervical cancer, and our findings revealed that patients in group 1 (OALO) who overexpressed ATP5H had a favorable prognosis, while those in group 4 (LASONH) showing low expression levels of ATP5H and SCP3 and overexpression of NANOG had the worst prognosis. Patients in group 2 (LASIM), showing low expression levels of ATP5H and SCP3, and those in group 3 (LASNIM) with low expression levels of ATP5H, SCP3, and NANOG had intermediate clinical outcomes. Notably, our results demonstrated higher accuracy with a higher C-index for recurrence than with traditionally used prognostic factors, such as FIGO staging and LN metastasis. In addition, clinicopathological characteristics, such as FIGO stage, tumor size, LVSI, LN metastasis, and response to CCRT, gradually worsened as the groups changed from 1 to 4. Our results indicate that AI technology can facilitate the development of personalized treatment strategies for patients with cervical cancer based on our knowledge of the molecular mechanism of cervical cancer growth [26, 27]. For example, for clinical implementation, if a young female diagnosed with cervical cancer undergoes radical hysterectomy at an early stage and is not recommended to undergo adjuvant therapy but is diagnosed to have group 4 cancer according to the molecular classification system, gynecologists may choose adjuvant therapy, chemotherapy, or frequent follow-ups, because this patient is at a high risk of LVSI, LN metastasis, and recurrence. In addition, patients who were at high risk of relapse, such as those with positive LNs, positive resec-



Variable	C-index (95% CI)	P-value
New classification+ FIGO stage (>2A) + Tumor size (>4cm) + LNM	0.728 (0.653-0.803)	<.001
New classification	0.683 (0.612-0.755)	<.001
FIGO stage (>2A)	0.608 (0.559-0.657)	<.001
FIGO stage (2A) + LNM + Tumor size (>4cm)	0.592 (0.505-0.679)	0.073
LNM	0.575 (0.525-0.625)	<.001
Tumor size (>4cm)	0.570 (0.520-0.621)	0.016
Grade (Poor)	0.527 (0.473-0.580)	0.105
Age (>50)	0.493 (0.445-0.541)	0.562

Figure 2. Assessment of an integrative clinico-molecular model for cervical cancer prognosis on the test cohort. A. Comparison of concordance index (C-index) for clinical factor, molecular classification and combination of clinical information and molecular classification. B. Time - dependent area under the ROC curve (AUC) of disease free survival. LNM, Lymph node metastasis.

tion margins, or parametrial invasion according to pathology reports, underwent CCRT after surgical resection. However, when a patient is diagnosed to have molecular group 4 cancer, which shows a poor response to CCRT, gynecologists could select additional systemic chemotherapy or therapy using anti-angiogenesis inhibitors to improve survival, guide personalized clinical treatment, and provide a strong treatment plan for patients with poor prognosis to improve the patient survival rate [28]. Notably, due to the development of the molecular classification system, the combination of clinical factors with the molecular classification significantly increased the predictive power of prognosis.

Recent advancements in transcriptomic, proteomic, and AI technologies have enabled the identification of 10 biomarkers, including ATP5H, SCP3, pERK, NANOG, PTEN, p16, CRY1, TRPV1, and FOXO1, that exhibit significantly different expression levels in cervical cancer tissues and nonadjacent normal epithelial tissues. These biomarkers are associated with the prognosis of cervical cancer and the response to chemotherapy. Using a Random Forest model, ATP5H, SCP3, and NANOG were identified to be the most critical biomarkers for molecular classification, with the potential to predict the survival of patients with cervical cancer. Previous studies have revealed that the levels of ATP5H, the D subunit of mitochondrial ATP synthase, are lower in cervical cancer specimens than in healthy controls, and this difference in ATP5H levels is associated with cervical cancer progression and poor prognosis [12, 13]. SCP3, a synaptonemal complex protein involved in DNA binding and chromosome pairing during meiosis, is associated with aggressive disease and poor prognosis in patients with cervical cancer [13, 29]. Moreover, NANOG functions as a key cancer stem cell transcription factor that has been found to be significantly overexpressed in cervical cancer specimens and is correlated with poor response to chemoradiotherapy and poor PFS in patients with cervical cancer [13, 16, 18, 19, 30-32]. Additionally, beyond the analysis of the clinicopathological characteristics associated with these biomarkers, recent studies have examined the signaling pathways involved in cervical cancer progression and therapeutic resistance to provide novel molecular classifications. For example, Cho *et al*., reported that loss of ATP5H leads to mitochondrial reprogramming, which promotes AKT activation and resistance to therapy, whereas SCP3 overexpression promotes AKT-mediated tumorigenesis [12, 29]. In addition, Noh *et al*., found that activation of the AKT signaling pathway promotes the expression of NANOG, leading to the acquisition of cancer stem cell-like properties and immune evasion in cancer cells [32]. Furthermore, AKT activation promotes the cyclin D-CDK4/6 pathway, leading to the overexpression of NANOG and SCP3, which leads to sensitization of therapy-refractory cancer [13]. Interestingly, these biomarkers are closely related to the AKT pathway, which plays a critical role in the initiation and progression of cervical cancer [33]. HPV-induced infection, followed by the expression of E6 or E7, activates the AKT signaling pathway, leading to malignancy initiation, cell proliferation, metastasis, and drug resistance, which are the primary pathways in cervical cancer [34]. Therefore, the selection of these biomarkers, which are closely related to the AKT pathway, is relevant for novel molecular classification and detection of resistance to therapy, as these biomarkers are involved in the primary pathway of cervical carcinogenesis. Further research on these biomarkers and their interactions with the AKT pathway could lead to a better understanding of cervical cancer initiation and progression and potentially improve the clinical management of cervical cancer with a high C-index. These findings provide a solid foundation for the development and validation of a model for predicting cervical cancer prognosis and offer valuable clinical insights for improving patient outcomes.

Limitations of the study

The notable strength of our study is demonstrated by the linear model's performance, which closely aligns with that of the XGBoost algorithm, as detailed in Supplementary Figure 4. Despite XGBoost's recognition for its advanced capabilities, out linear model not only competes effectively in terms of accuracy but also excels in providing transparency and interpretability. These qualities are indispensable for informed clinical decision-making within the realm of precision medicine. However, it is imperative to acknowledge the study's limitations, such as its modest sample size and the confines of a single-center retrospective design, which necessitate a prudent interpretation of the results. To ascertain the clinical validity and reliability of our findings, further validation in a broader, multicenter cohort is essential. Pursuing this line of inquiry has the potential to refine classification methodologies, thereby enhancing patient outcomes and shaping individualized treatment modalities in cervical cancer management.

Conclusions

In conclusion, our investigation elucidates that an integrative molecular classification framework, devised via the analysis of surgical specimen and augmented by an artificial intelligence methodologist, possesses the capability to prognosticate clinical outcomes, discern patients with heightened recurrence risk, and appraise treatment efficacy in cervical cancer. Hence, this classification paradigm could by instrumental in refining personalized therapeutic stratification for individuals afflicted with cervical cancer, thereby advancing precision oncology.

Acknowledgements

This study was supported by a National Research Foundation of Korea (NRF) grant, funded by the Korean government (MIST; grant nos. NRF-2020R1A2C2004782 and NRF-2022R1-F1A1071806). This study was also supported by a faculty research grant from the Yonsei University College of Medicine (grant no. 6-2020-0226), and the Inje University research grant (grant no. 20230084).

Disclosure of conflict of interest

None.

Address correspondence to: Dr. Hanbyoul Cho, Department of Obstetrics and Gynecology, Gangnam Severance Hospital, Yonsei University College of Medicine, 211 Eonju-Ro, Gangnam-Gu, Seoul 06273, Republic of Korea. Tel: +82-2-2019-3430; ORCID: 0000-0002-6177-1648; Fax: +82-2-3462-8209; E-mail: hanbyoul@yuhs.ac

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A and Bray F. Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021; 71: 209-249.

[2] White EJ, Nacke M, Akeman E, Cannon MJ, Mayeli A, Touthang J, Zoubi OA, McDermott TJ, Kirlic N, Santiago J, Kuplicki R, Bodurka J, Paulus MP, Craske MG, Wolitzky-Taylor K, Abelson J, Martell C, Clausen A, Stewart JL and Aupperle RL. P300 amplitude during a monetary incentive delay task predicts future therapy completion in individuals with major depressive disorder. J Affect Disord 2021; 295: 873-882.

[3] Tang X, Guo C, Liu S, Guo J, Hua K and Qiu J. A novel prognostic nomogram utilizing the 2018 FIGO staging system for cervical cancer: a large multicenter study. Int J Gynaecol Obstet 2021; 155: 86-94.

[4] Wang E, Panelli M and Marincola F. Genomic analysis of cancer. Princ Pract Oncol 2003; 17: 1-16.

[5] Williams VM, Filippova M, Soto U and Duerksen-Hughes PJ. HPV-DNA integration and carcinogenesis: putative roles for inflammation and oxidative stress. Future Virol 2011; 6: 45-57.

[6] Ravdin PM, Siminoff LA, Davis GJ, Mercer MB, Hewlett J, Gerson N and Parker HL. Computer program to assist in making decisions about adjuvant therapy for women with early breast cancer. J Clin Oncol 2001; 19: 980-991.

[7] Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJ and Yang PC. A five-gene signature and clinical outcome in non-small-cell lung cancer. N Engl J Med 2007; 356: 11-20.

[8] Wang HY, Sun BY, Zhu ZH, Chang ET, To KF, Hwang JS, Jiang H, Kam MK, Chen G, Cheah SL, Lee M, Liu ZW, Chen J, Zhang JX, Zhang HZ, He JH, Chen FL, Zhu XD, Huang MY, Liao DZ, Fu J, Shao Q, Cai MB, Du ZM, Yan LX, Hu CF, Ng HK, Wee JT, Qian CN, Liu Q, Ernberg I, Ye W, Adami HO, Chan AT, Zeng YX and Shao JY. Eight-signature classifier for prediction of nasopharyngeal [corrected] carcinoma survival. J Clin Oncol 2011; 29: 4516-4525.

[9] Muñoz N, Bosch FX, de Sanjosé S, Herrero R, Castellsagué X, Shah KV, Snijders PJ and Meijer CJ; International Agency for Research on Cancer Multicenter Cervical Cancer Study Group. Epidemiologic classification of human papillomavirus types associated with cervical cancer. N Engl J Med 2003; 348: 518-527.

[10] Bhatla N and Denny L. FIGO cancer report 2018. Int J Gynaecol Obstet 2018; 143 Suppl 2: 2-3.

[11] Talia KL, Oliva E, Rabban JT, Singh N, Stolnicu S and McCluggage WG. Grading of endocervical adenocarcinomas: review of the literature and recommendations from the International Society of Gynecological Pathologists. Int J Gynecol Pathol 2021; 40 Suppl 1: S66-S74.

[12] Song KH, Kim JH, Lee YH, Bae HC, Lee HJ, Woo SR, Oh SJ, Lee KM, Yee C, Kim BW, Cho H, Chung EJ, Chung JY, Hewitt SM, Chung TW, Ha KT, Bae YK, Mao CP, Yang A, Wu TC and Kim TW. Mitochondrial reprogramming via ATP5H loss promotes multimodal cancer therapy resistance. J Clin Invest 2018; 128: 4098-4114.

[13] Oh SJ, Cho H, Kim S, Noh KH, Song KH, Lee HJ, Woo SR, Kim S, Choi CH, Chung JY, Hewitt SM, Kim JH, Baek S, Lee KM, Yee C, Park HC and Kim TW. Targeting Cyclin D-CDK4/6 sensitizes immune-refractory cancer by blocking the SCP3-NANOG axis. Cancer Res 2018; 78: 2638-2653.

[14] Luna AJ, Sterk RT, Griego-Fisher AM, Chung JY, Berggren KL, Bondu V, Barraza-Flores P, Cowan AT, Gan GN, Yilmaz E, Cho H, Kim JH, Hewitt SM, Bauman JE and Ozbun MA. MEK/ERK signaling is a critical regulator of high-risk human papillomavirus oncogene expression revealing therapeutic targets for HPV-induced tumors. PLoS Pathog 2021; 17: e1009216.

[15] Han GH, Chay DB, Nam S, Cho H, Chung JY and Kim JH. The combination of transient receptor potential vanilloid type 1 (TRPV1) and phosphatase and tension homolog (PTEN) is an effective prognostic biomarker in cervical cancer. Int J Gynecol Pathol 2021; 40: 214-223.

[16] Han GH, Kim J, Yun H, Cho H, Chung JY, Kim JH and Hewitt SM. CRY1 regulates chemoresistance in association with NANOG by inhibiting apoptosis via STAT3 pathway in patients with cervical cancer. Cancer Genomics Proteomics 2021; 18: 699-713.

[17] Chay DB, Han GH, Nam S, Cho H, Chung JY and Hewitt SM. Forkhead box protein O1 (FOXO1) and paired box gene 3 (PAX3) overexpression is associated with poor prognosis in patients with cervical cancer. Int J Clin Oncol 2019; 24: 1429-1439.

[18] Kim S, Cho H, Hong SO, Oh SJ, Lee HJ, Cho E, Woo SR, Song JS, Chung JY, Son SW, Yoon SM, Jeon YM, Jeon S, Yee C, Lee KM, Hewitt SM, Kim JH, Song KH and Kim TW. LC3B upregulation by NANOG promotes immune resistance and stem-like property through hyperactivation of EGFR signaling in immune-refractory tumor cells. Autophagy 2021; 17: 1978-1997.

[19] Song KH, Oh SJ, Kim S, Cho H, Lee HJ, Song JS, Chung JY, Cho E, Lee J, Jeon S, Yee C, Lee KM, Hewitt SM, Kim JH, Woo SR and Kim TW. HSP90A inhibition promotes anti-tumor immunity by reversing multi-modal resistance and stem-like property of immune-refractory tumors. Nat Commun 2020; 11: 562.

[20] Oh SJ, Lim JY, Son MK, Ahn JH, Song KH, Lee HJ, Kim S, Cho EH, Chung JY, Cho H, Kim H, Kim JH, Park J, Choi J, Hwang SW and Kim TW. TRPV1 inhibition overcomes cisplatin resistance by blocking autophagy-mediated hyperactivation of EGFR signaling pathway. Nat Commun 2023; 14: 2691.

[21] Kim BW, Cho H, Chung JY, Conway C, Ylaya K, Kim JH and Hewitt SM. Prognostic assessment of hypoxia and metabolic markers in cervical cancer using automated digital image analysis of immunohistochemistry. J Transl Med 2013; 11: 185.

[22] Helgason H, Eiriksdottir T, Ulfarsson MO, Choudhary A, Lund SH, Ivarsdottir EV, Hjorleifsson Eldjarn G, Einarsson G, Ferkingstad E, Moore KHS, Honarpour N, Liu T, Wang H, Hucko T, Sabatine MS, Morrow DA, Giugliano RP, Ostrowski SR, Pedersen OB, Bundgaard H, Erikstrup C, Arnar DO, Thorgeirsson G, Masson G, Magnusson OT, Saemundsdottir J, Gretars-dottir S, Steinthorsdottir V, Thorleifsson G, Helgadottir A, Sulem P, Thorsteinsdottir U, Holm H, Gudbjartsson D and Stefansson K. Evaluation of large-scale proteomics for prediction of cardiovascular events. JAMA 2023; 330: 725-735.

[23] Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Lønning PE and Børresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A 2001; 98: 10869-10874.

[24] Komor MA, Bosch LJ, Bounova G, Bolijn AS, Delis-van Diemen PM, Rausch C, Hoogstrate Y, Stubbs AP, de Jong M, Jenster G, van Grieken NC, Carvalho B, Wessels LF, Jimenez CR, Fijneman RJ and Meijer GA; NGS-ProToCol Consortium. Consensus molecular subtype classification of colorectal adenomas. J Pathol 2018; 246: 266-276.

[25] Jain S, Xu R, Prieto VG and Lee P. Molecular classification of soft tissue sarcomas and its clinical applications. Int J Clin Exp Pathol 2010; 3: 416-28.

[26] Obrzut B, Semczuk A, Naróg M, Obrzut M and Król P. Prognostic parameters for patients with cervical cancer FIGO stages IA2-IIB: a long-term follow-up. Oncology 2017; 93: 106-114.

[27] Ho CM, Chien TY, Huang SH, Wu CJ, Shih BY and Chang SC. Multivariate analysis of the prognostic factors and outcomes in early cervical cancer patients undergoing radical hysterectomy. Gynecol Oncol 2004; 93: 458-464.

[28] Hill EK. Updates in cervical cancer treatment. Clin Obstet Gynecol 2020; 63: 3-11.

[29] Cho H, Noh KH, Chung JY, Takikita M, Chung EJ, Kim BW, Hewitt SM, Kim TW and Kim JH. Synaptonemal complex protein 3 is a prognostic marker in cervical cancer. PLoS One 2014; 9: e98712.

[30] Yun H, Han GH, Kim J, Chung JY, Kim JH and Cho H. NANOG regulates epithelial-mesenchymal transition via AMPK/mTOR signalling pathway in ovarian cancer SKOV-3 and A2780 cells. J Cell Mol Med 2022; 26: 5277-5291.

[31] Song KH, Choi CH, Lee HJ, Oh SJ, Woo SR, Hong SO, Noh KH, Cho H, Chung EJ, Kim JH, Chung JY, Hewitt SM, Baek S, Lee KM, Yee C,

Son M, Mao CP, Wu TC and Kim TW. HDAC1 upregulation by NANOG promotes multidrug resistance and a stem-like phenotype in immune edited tumor cells. Cancer Res 2017; 77: 5039-5053.

[32] Noh KH, Kim BW, Song KH, Cho H, Lee YH, Kim JH, Chung JY, Kim JH, Hewitt SM, Seong SY, Mao CP, Wu TC and Kim TW. Nanog signaling in cancer promotes stem-like phenotype and immune evasion. J Clin Invest 2012; 122: 4077-4093.

[33] Gupta S, Kumar P and Das BC. HPV: molecular pathways and targets. Curr Probl Cancer 2018; 42: 161-174.

[34] Bhattacharjee R, Das SS, Biswal SS, Nath A, Das D, Basu A, Malik S, Kumar L, Kar S, Singh SK, Upadhye VJ, Iqbal D, Almojam S, Roychoudhury S, Ojha S, Ruokolainen J, Jha NK and Kesari KK. Mechanistic role of HPV-associated early proteins in cervical cancer: molecular pathways and targeted therapeutic strategies. Crit Rev Oncol Hematol 2022; 174: 103675.

Gene selection

The data set was divided into a training set (70%) and a test set (30%) with further stratification into derivation and validation sets within the training set. The molecular classification model was constructed using all the 27 protein expression levels. The new group was developed in the deviation set using a cox proportional hazards model with a lasso penalty by solving.

$$\overset{min}{\beta}\left(-\log(\beta) + \lambda \parallel \beta \parallel_1\right),$$

Where β is a vector of variable coefficients, the parameter λ >0 controls penalization strength (large λ correspond to greater penalization), and log(β) is the partial likelihood function of the Cox model. The parameter λ was chosen to minimize mean partial likelihood function of the Cox model. The parameter λ was chosen to minimize mean partial likelihood deviance in the hold-out set of ten-fold cross validation in the derivation set. The glmnet package in R was used to perform the fit. To ensure the coefficient for age, i.e βage would be non-zero, they were exempted from the penalization. This was done to prevent the model from trying to capture the effects of age on the endpoint using the gene variables. For the chosen penalization strength, the model had 7 other non-zero coefficient βk. Further to estimate the robustness of proteins selected by lasso-penalized Cox method, we performed bootstrapping. Using resampling with replacement, we sampled 100 different sets of 134 participants from deviation set in the recurrent population. In each seat, we trained lasso-penalized Cox models, for 20 different values of model size penalization parameter λ using all the gene measurement, age, where age was excluded from penalization (i.e forced into the model). For the λ closet to the λ used to derivation gene, we counted how often each gene was included in the 1000 different models.

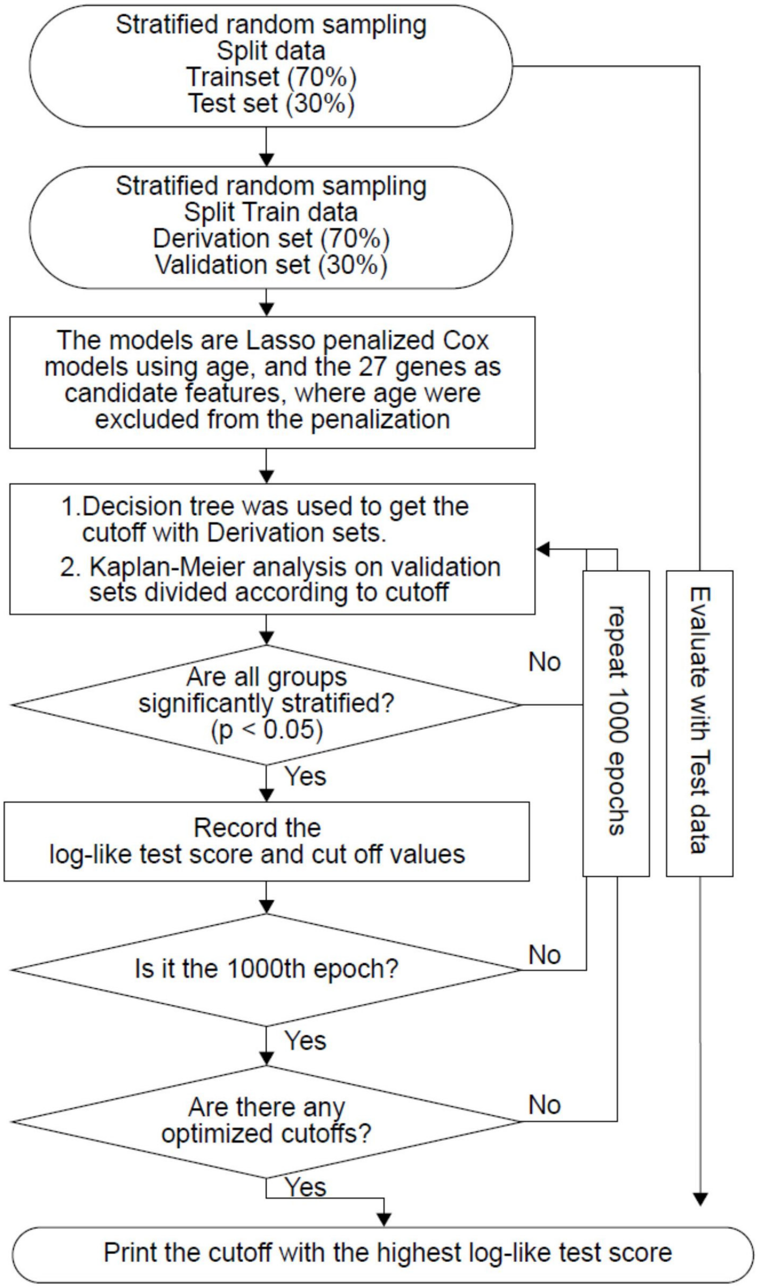
Development of molecular classification with

To tried non-linear methos to create a risk score. As a non-linear approach for creating risk score, we trained a Cox proportional hazard model with gradient boosted trees using XGBoost [29]. We used random search which cross validation to adjust the following tree hyperparameters: mx tree depth, learning rate, data instance subsampling, variable sampling, minimum child weight. To further prevent overfitting, we used early stopping to determine the number of boosting iterations. The non-linear and linear models were compared using five-fold cross-validation on the derivation set and we found the performance of both models to be similar. Risk score were calculating using XGBoost, rBayesianOptimization, surv XGBoost and careEnsemble packages in R. A decision tree algorithm was used to obtain the optimal cut-off point. The C4.5 algorithm is one of the classification algorithms used to create decision trees. The Gini index is one of the crucial metrics used in the C4.5 algorithm as a criterion for node splitting. Decision trees are used to divide data based on specific attributes and classify them into different classes. We selected the cut-off point that minimizes the Gini index by dividing the train data of the continuous variable at all possible split points. The Gini index is employed to measure the impurity of a particular node, where impurity refers to how well the data in the node is mixed with different classes. In other words, a higher Gini index indicates that the data in the node is spread across multiple classes, leading to higher uncertainty.

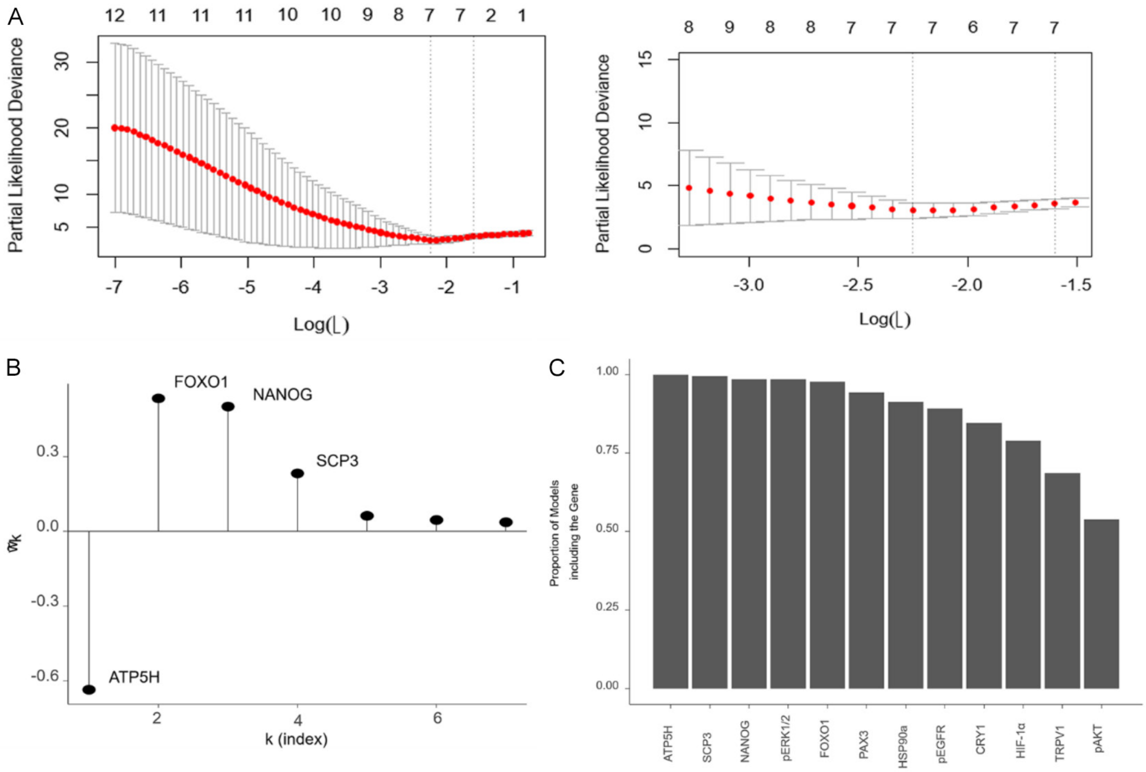
$$\text{Gini index} = 1 - \sum_{i=1}^{c} p_i^2 = 1 - \left[p(Y = 0)\right]^2 - \left[p(Y = 1)\right]^2$$

c = class number, Y = overall survival or progression-free survival.

The log-rank test and the Kaplan-Meier analysis was performed to ensure that all clusters were significantly stratified. During the training, accuracy increases with the epoch flow up from 1 to 1000, and best optimal cut off with the highest log-likelihood test score was selected for cut off value.


Stratified random sampling
Split data
Trainset (70%)
Test set (30%)
Stratified random sampling
Split Train data
Derivation set (70%)
Validation set (30%)
The models are Lasso penalized Cox models using age, and the 27 genes as candidate features, where age were excluded from the penalization
1.Decision tree was used to get the cutoff with Derivation sets.
2. Kaplan-Meier analysis on validation sets divided according to cutoff
Are all groups significantly stratified? (p < 0.05)
No
Yes
Record the log-like test score and cut off values
Is it the 1000th epoch?
No
Yes
Are there any optimized cutoffs?
No
Yes
repeat 1000 epochs
Evaluate with Test data
Print the cutoff with the highest log-like test score

Supplementary Figure 1. Overall data processing algorithm and detailed methods.

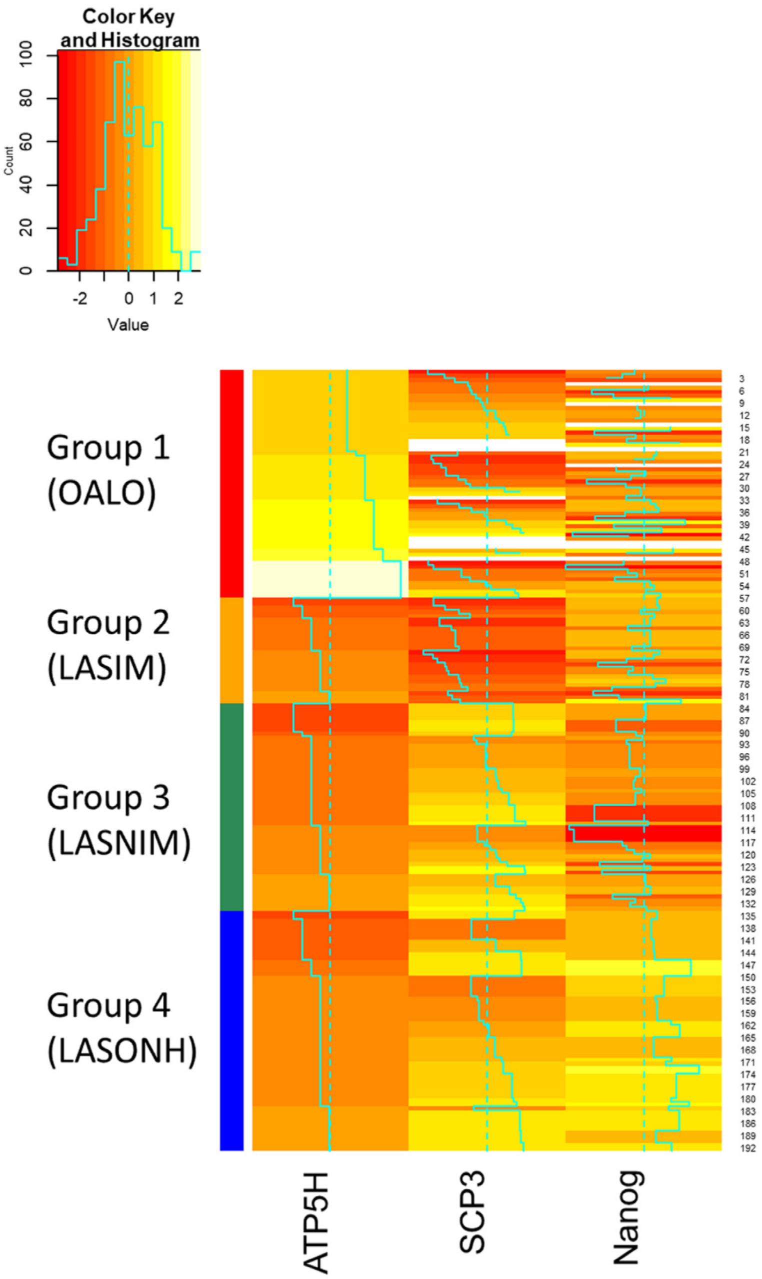


Supplementary Figure 2. Stratified prognostic biomarker assessment process. (A) Cross-validation for selecting penalization strength for fitting gene DFS risk prediction models. The figure demonstrates how the penalization strength λ for fitting the Risk score using all the Gene measurements, age, where age is excluded from the penalization, is selected based on the partial likelihood deviance (PLD) in ten-fold cross-validation. The red dots represent the mean PLD for each λ that was used and the bars represent one standard deviation in each direction for the cross-validations. The top x-axis shows how the number of non-zero coefficients changes with the penalization strength. For each panel, the vertical dotted line to the left indicates the λ that gave the lowest mean PLD; this is the penalization strength λm that was used to develop the DFS risk score. The vertical line to the right indicates the highest λ that gave a mean PLD within one standard deviation from the PLD corresponding to λm. (A) shows the PLDs for all 100 tested λs while (B) shows in close up the PLDs for the λs closest to λm. (B) Risk score predicting disease free survival weights and relationship with risk scores with derivation set. (C) shows the frequency of inclusion of the different gene in multiple DFS risk models. The models are Lasso penalized Cox models using age, and the 27 genes as candidate features, where age was excluded from the penalization, trained on 1,000 different resamplings of the derivation data. Out of the twenty tried penalization strengths, these figures show results for the λ closest to λm used to train DFS risk score. Shown is the proportion of models that include the 12 genes that are included in at least 50% of the models; the labels on the x-axis are the gene names. ATP5H is the only protein that is included in every model and SCP3, NANOG, pERK1/2, and FOXO1 was other gene included in more than 95% of the models.

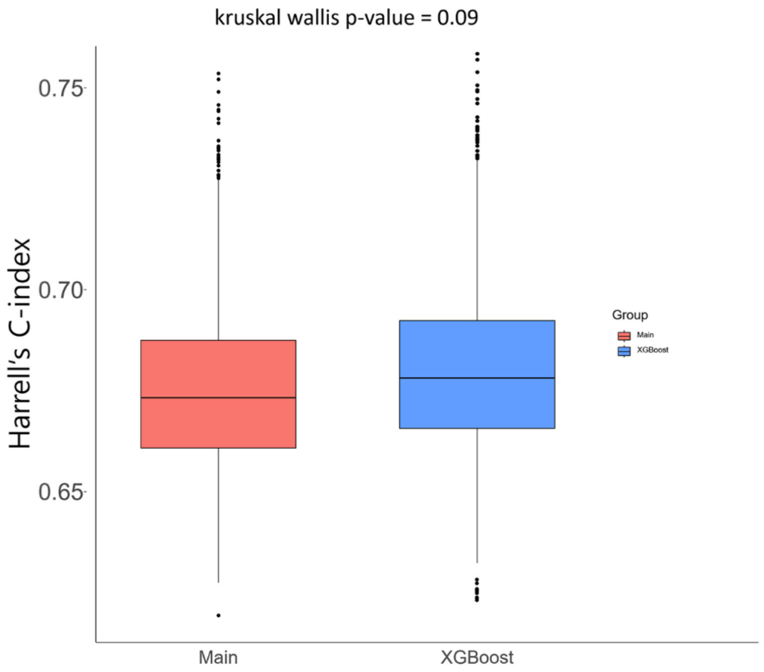
Supplementary Table 1. Probes and weights for DFS risk score trained including age covariates

k (index)**	Gene Name	wk (weight)*	Weight for standardizing with training set: w'k	Standard deviation of standardizing probe value: sk	Average standardizing probe value: mk	Proportion of Absolute Standardized Weight of Total Norm of Weights	HR (per SD)	*p*-value
1	ATP5H	-0.31550225	-0.913616025	2.895751223	4.037558685	63.44255676	0.20 (0.14-0.29)	< 0.001
2	FOXO1	0.019658418	0.767939434	39.06415233	217.5776398	53.32660525	1.00 (0.78-1.27)	0.99
3	NANOG	0.015793919	0.721353092	45.67283675	173.2254902	50.09159565	1.50 (1.24-1.82)	< 0.001
4	SCP3	0.124533849	0.334557866	2.686481366	7.066666667	23.23208634	2.04 (1.64-2.54)	< 0.001
5	pERK1/2	0.002157925	0.090910846	42.12881926	48.04166667	6.31295462	0.86 (0.69-1.07)	0.19
6	pAKT	0.001190521	0.066616001	55.9553346	223.3680982	4.625892425	1.61 (1.28-2.03)	< 0.001
7	LC3B	0.00085308	0.05371537	62.96639609	160.5939394	3.730057591	1.44 (1.16-1.77)	< 0.001

*The protein score is given by the formula below, where xk is the raw probe measurement for probe with index k (see Methods); **The weights wk are sorted based on the decreased order for absolute weights corresponding to standardization of probe measurements in the training set (w'k); note that wk = w'k/sk. All results reported in this table are obtained from the derivation set N = 134. Column descriptions: SeqId: Identifier of SOMAScan aptamer; Gene Name: The gene name corresponding to protein target; Target Name: Short-hand name of the targeted protein; Target Full Name: Long-hand name for the targeted protain; UniProt: The UniProt identifier for the targeted protein.



Supplementary Figure 3. The heatmap of molecular classification model of cervical cancer.



Supplementary Figure 4. The figure shows the Harrell's C-indices were calculated by bootsraping performed in the Test data set (N = 89). Using resampling with replacement we sampled 100 different sets of participants from the Test set in the primary event population. We performed the 1000 different models. These results demonstrate no obvious benefit from using the non-linear gradient boosted tree cox model or the simpler penalized linear cox model over Decision Tree algorithm.