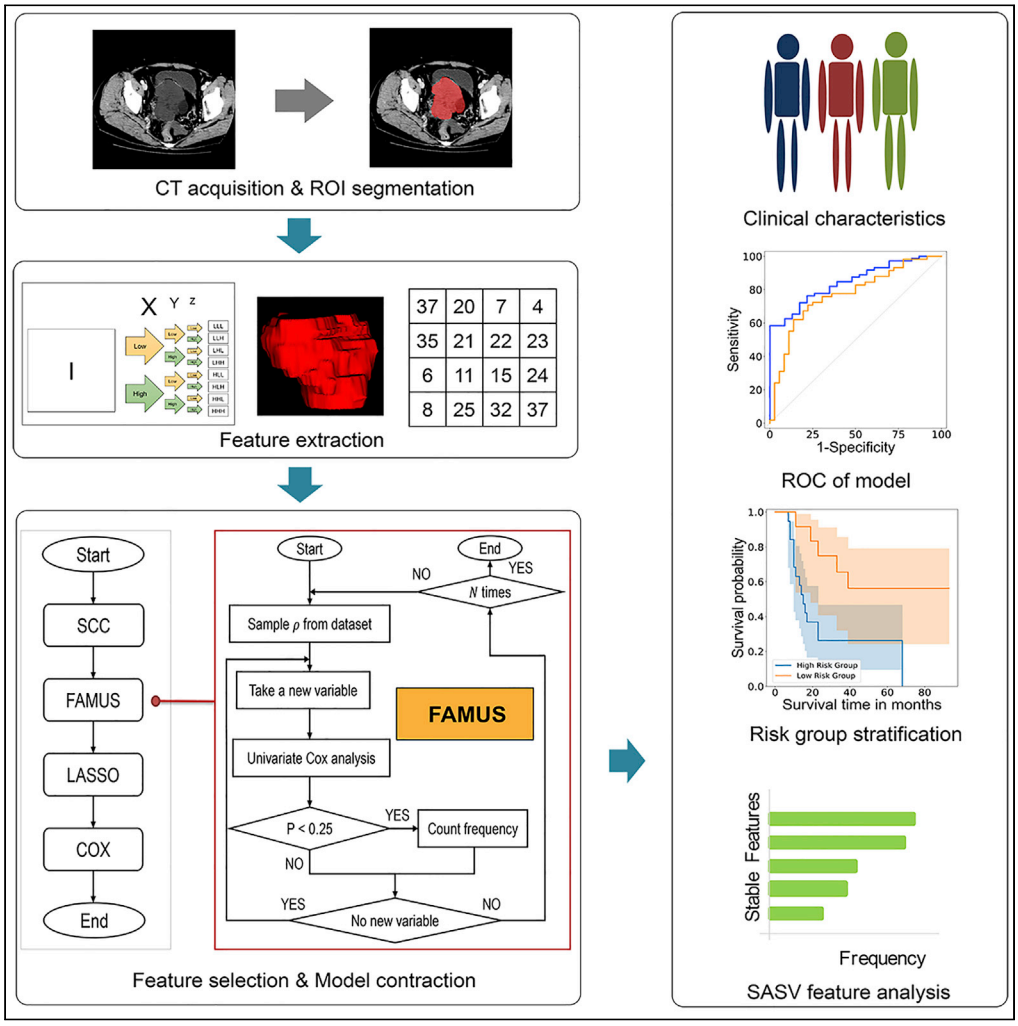
iScience

CellPress
OPEN ACCESS

Article

Development of survival predictors for high-grade serous ovarian cancer based on stable radiomic features from computed tomography images



Jiaqi Hu, Zhiwu Wang, Ruocheng Zuo, ..., Chunhui Zhao, Pengyuan Liu, Yan Lu

pyliu@zju.edu.cn (P.L.)
yanlu76@zju.edu.cn (Y.L.)

Highlights

Frequency Appearance in Multiple Univariate preScreening (FAMUS) identifies stable and task-relevant radiomic features from computed tomography (CT) images

Radiomics-based signatures are highly predictive of the clinical outcome of high-grade serous ovarian cancer (HGSOC)

FAMUS improves the prognostic performance of radiomics-based prediction models

Developed radiomic models can help clinicians tailor treatment plans for HGSOC

Hu et al., iScience 25, 104628
July 15, 2022
https://doi.org/10.1016/j.isci.2022.104628

iScience

CellPress
OPEN ACCESS

Article

Development of survival predictors for high-grade serous ovarian cancer based on stable radiomic features from computed tomography images

Jiaqi Hu,[1,8] Zhiwu Wang,[2,8] Ruocheng Zuo,[1,8] Chengcai Zheng,[1,3] Bingjian Lu,[1,7] Xiaodong Cheng,[1,7] Weiguo Lu,[4,7] Chunhui Zhao,[6] Pengyuan Liu,[5,7,*] and Yan Lu[1,7,9,*]

SUMMARY

Less than 35% of advanced patients with high-grade serous ovarian cancer (HGSOC) survive for 5 years after diagnosis. Here, we developed radiomics-based models to predict HGSOC clinical outcomes using preoperative contrast-enhanced computed tomography (CECT) images. 891 radiomics features were extracted between primary, metastatic, or lymphatic lesions from preoperative venous phase CECT images of 217 patients with HGSOC. A heuristic method, Frequency Appearance in Multiple Univariate preScreening (FAMUS), was proposed to identify stable and task-relevant radiomic features. Using FAMUS, we constructed predictive models of overall survival and disease-free survival in patients with HGSOC based on these stable radiomic features. According to their CT images, patients with HGSOC can be accurately stratified into high-risk or low-risk groups for cancer-related death within 2-6 years or for likely recurrence within 1-5 years. These radiomic models provide convincing and reliable non-invasive markers for individualized prognostic evaluation and clinical decision-making for patients with HGSOC.

INTRODUCTION

Ovarian cancer is a leading global cause of cancer-related deaths in women, accounting for 4.7% of all female cancer deaths (Sung et al., 2021), and showing a rising trend in mortality (Chen et al., 2016). High-grade serous ovarian cancer (HGSOC) is the most frequently occurring gynecological malignancy, accounting for 70-80% of ovarian cancer deaths (Bowtell et al., 2015). The 5-year survival rate of patients with HGSOC is only 34% (Vang et al., 2009), with relapse being the main factor responsible for such a high mortality rate (Vaughan et al., 2011). The identification of new biomarkers to predict clinical outcomes of ovarian cancer, which will facilitate the timely inclusion of patients with poor prognosis into personalized treatment strategies or clinical trials, is of high urgency.

Contrast-enhanced computed tomography (CECT) plays an essential role in HGSOC diagnosis and treatment assessment, providing a low-cost and non-invasive method to extract prognostic information (Nougaret et al., 2017). Radiomics has been proposed to explore the correlation between medical images and underlying genetic characteristics, enabling the generation of a comprehensive overview of the spatiotemporal heterogeneity of tumors. The use and applications of radiomics are similarly expanding in many spheres of biology (Aerts et al., 2014), with varied studies showing applications for radiomics toward improvements in chemoradiotherapy effectiveness (Xie et al., 2019), the classification of molecular characteristics (Rathore et al., 2018), or in distant metastasis prediction (Wu et al., 2017), and so forth.

Feature selection is a crucial step for high-dimensional data analysis in radiomics. By removing irrelevant and "noisy" factors, feature selection makes the overall analysis more manageable, efficient, and productive (Yu and Liu, 2003). The sensitivity of feature selection results to sample variation has been a recent focus of discussion (Kalousis et al., 2007). When applying feature selection aimed at a specific targeted discovery, the high stability of the feature selection result is one of the key focus areas. Though various routinely used radiomic feature selection methods have a good performance, improving the stability of features against sample variation is still desirable to further increase the robustness and efficacy of such radiomic models.

[1]Zhejiang Provincial Key Laboratory of Precision Diagnosis and Therapy for Major Gynecological Diseases, Department of Gynecologic Oncology, Women's Hospital and Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310006, China

[2]Department of Chemoradiotherapy, Tangshan People's Hospital, Tangshan, Hebei 063000, China

[3]Department of Critical Care Medicine, the First Affiliated Hospital, Wenzhou Medical University, Wenzhou, Zhejiang 325000, China

[4]Women's Reproductive Health Key Laboratory of Zhejiang Province, Women's Hospital and Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310006, China

[5]Key Laboratory of Precision Medicine in Diagnosis and Monitoring Research of Zhejiang Province, Sir Run Run Shaw Hospital and Institute of Translational Medicine, Zhejiang University School of Medicine, Hangzhou, Zhejiang 310016, China

[6]State Key Laboratory of Industrial Control Technology, College of Control Science and Engineering, Zhejiang University, Hangzhou, Zhejiang 310007, China

[7]Cancer Center, Zhejiang University, Hangzhou, Zhejiang 310013, China

[8]These authors contributed equally

[9]Lead contact

*Correspondence: pyliu@zju.edu.cn (P.L.), yanlu76@zju.edu.cn (Y.L.)

https://doi.org/10.1016/j.isci.2022.104628

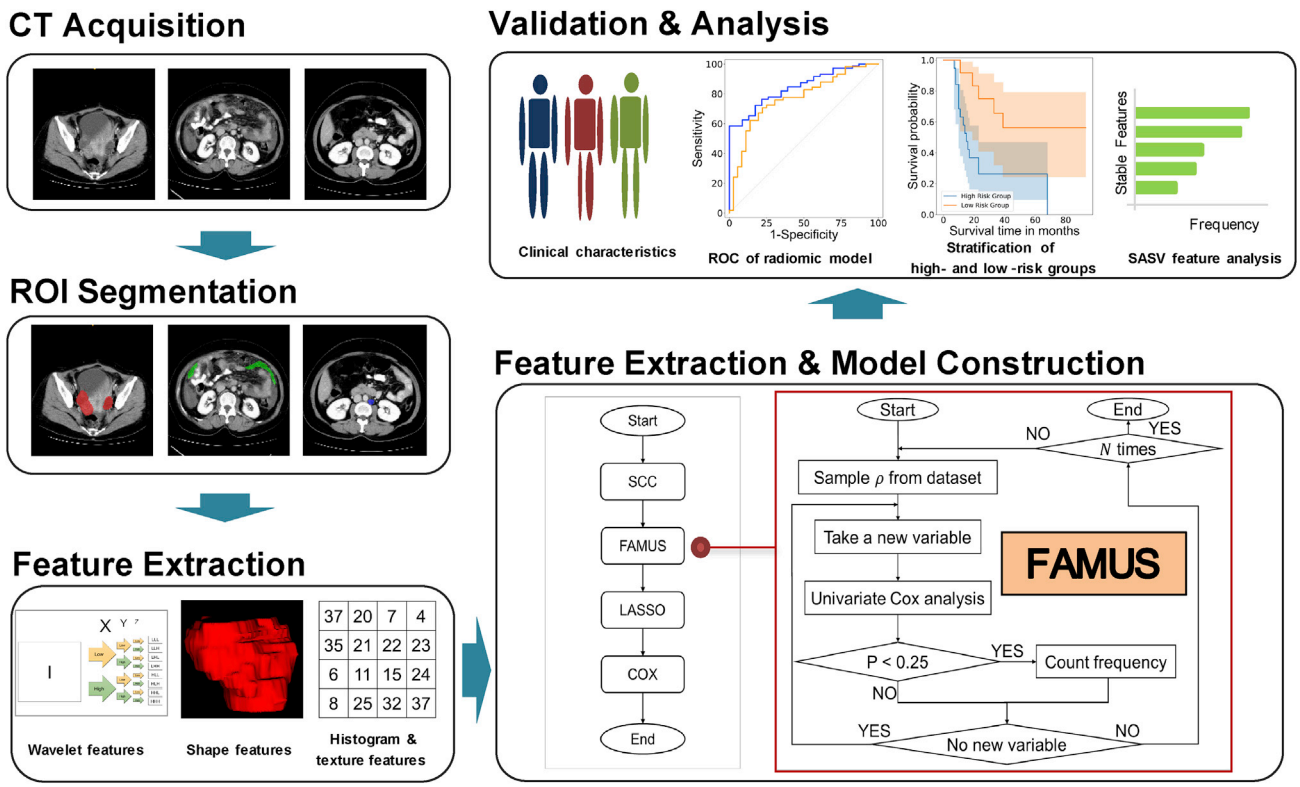


Figure 1. Flowchart of building prognosis prediction models using CT images

Briefly, the steps of establishing prognosis prediction models include CT acquisition, ROI segmentation, preliminary feature selection, detection of stable features by FAMUS, model construction, data analysis, and validation. FAMUS represents Frequency Appearance in Multiple Univariate preScreening, a heuristic algorithm to select stable features for constructing prognosis prediction models.

In this work, we propose a heuristic method, termed Frequency Appearance in Multiple Univariate pre-Screening (FAMUS), to identify radiomic features with stability against sample variation (SASV). Using FAMUS, we construct and validate prognostic models for predicting overall survival (OS) and disease-free survival (DFS) in patients with HGSOC based on radiomic features extracted from venous phase enhanced-CT images of preoperative patients (Figure 1). Two sets of three most robust, non-redundant, and predictive SASV features were selected to develop prognostic models for individualized, preoperative evaluation of OS and DFS in patients with HGSOC, respectively. The performance of these two radiomic models was evaluated using both internal and external validation cohorts. Additionally, two nomograms were constructed in conjunction with the corresponding radiomic signatures. These are proposed as a low-cost and non-invasive means for predicting the risk for HGSOC-related death or relapse.

RESULTS

Clinical characteristics and outcomes of patients

A total of 217 patients with HGSOC from the Women's Hospital of Zhejiang University School of Medicine and the First Affiliated Hospital of Wenzhou Medical University were enrolled in this study. Table 1 depicts the clinical characteristics and outcomes of these patient samples in this study. There were no significant differences between the two hospitals in age ($p = 0.45$), stage ($p = 0.422$), metastasis lesions ($p = 0.346$), or patient vital status ($p = 0.88$), but there was significant difference in lymphoid lesions ($p = 0.003$). Between all patients, the median (IQR) death and recurrence time were 30.0 months (17.0–46.0 months) and 17.9 months (13.0–26.7 months), respectively. The median (IQR) follow-up time for OS and DFS was 56.0 months (34.5–71 months) and 56.0 months (35.0–70.0 months) in the censored patients, respectively. There was no significant difference in OS between the training cohort and two validation cohorts, but there was a marginal difference in DFS (Figure S1).

Table 1. Demographic and clinical characteristics and outcomes of patients in three cohorts

Characteristics	Training cohort	In-valid cohort	Ex-valid cohort	P-value[a]
Age, years (median, range)	50 (20–73)	51.5 (18–73)	55 (32–68)	0.450
Clinical stage				0.422
I	13	14	5	
II	28	13	2	
III	49	56	16	
IV	3	4	1	
Metastatic lesion				0.346
Yes	54	47	14	
No	41	43	18	
Lymphoid lesion				0.003
Yes	68	83	18	
No	27	7	14	
Death				0.880
Yes	37	15	10	
No	58	75	22	
Recurrence				0.068
Yes	49	31	20	
No	46	59	12	
Follow-up in death (month), Median (IQR)	35(18, 50)	18(12.5, 27.5)	26(18.5, 48.25)	/
Follow-up in recurrence (month), Median (IQR)	24(15, 36)	15(11.5, 21)	14.5(10, 23)	/
Follow-up in censored patients for OS (month), Median (IQR)	71.5(62, 92)	38(26.5, 58)	42(35.25, 60.75)	/
Follow-up in censored patients for DFS, Median (IQR)	70.5(62, 86)	38(28, 56.5)	41(36, 63.75)	/

[a]The differences in clinical characteristics in the three cohorts were compared by using the Kruskal-Wallis test or Chi-square test.

Extraction and selection of radiomic features

A total of 851 radiomic features were extracted for each patient based on the regions of interest (ROI) of CT images, including 14 shape features, 19 histogram features, 23 Gy-level co-occurrence matrix features, 16 Gy-level size zone matrix features, 16 Gy-level run length matrix features, five neighborhood gray-tone difference matrix features, 14 Gy-level dependence matrix features and 744 wavelet-based features (Table S1 and Figure S2). Owing to the high-dimensional feature size, the feature selection procedure was performed to remove irrelevant and redundant information which may greatly degrade the prognostic performance of the learning algorithms. It included three main steps, as shown in Figure 1. Firstly, redundant features were eliminated using the Spearman correlation coefficient (SCC). When SCC was >0.8 between two features with p value <0.05, the one with larger p value, as calculated by univariate Cox regression, was regarded as redundant and removed. SCC analysis revealed that a total of 130 and 93 non-redundant features were significantly correlated with OS and DFS, respectively.

Secondly, the stability of the SCC-filtered features against sample variation was further evaluated. Small changes in the datasets often result in a different subset of selected features. This decreases the reliability of the feature selection results. To identify features with good stability against sample variation (SASV), a method of Frequency Appearance in Multiple Univariate preScreening (FAMUS) was proposed (Figure 1). In this method, a subset of samples was drawn randomly from the training dataset, and then the univariate Cox regression model was used for selecting candidate features. A p value <0.25 instead of <0.05 was taken as a threshold for feature prescreening as suggested (Grant et al., 2019) to avoid important covariate variables being dropped from the model owing to stochastic variability. To obtain a more versatile subset of data, the data were separately sampled from failing and censored patients at a sampling rate ρ. A total of N samplings were performed, and the frequency of each feature appearing in the selection process was accounted. Through our Monte Carlo simulation analysis, we determined that $N = 2000$ and $\rho = 1/6$ were appropriate for our study (Table S2 and Figure S3). The frequencies of SASV features are expected

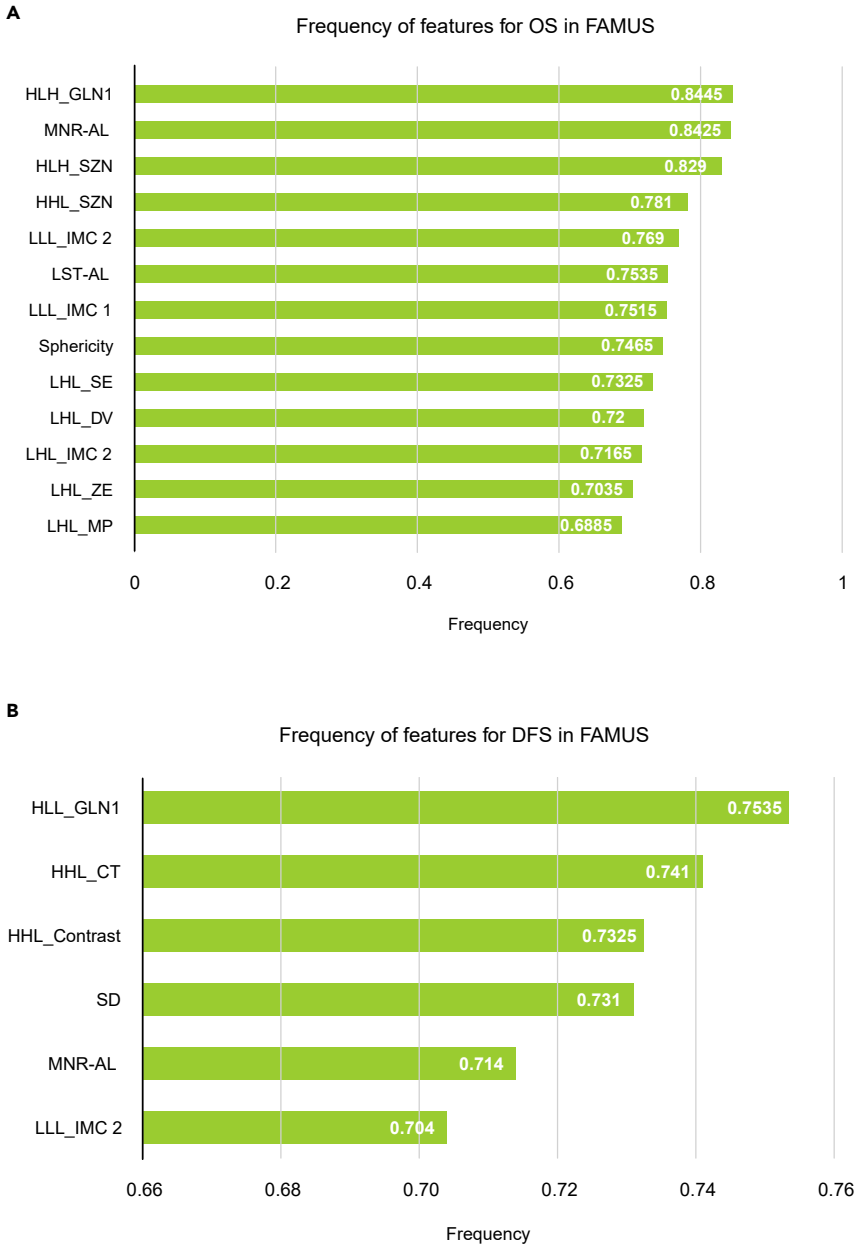


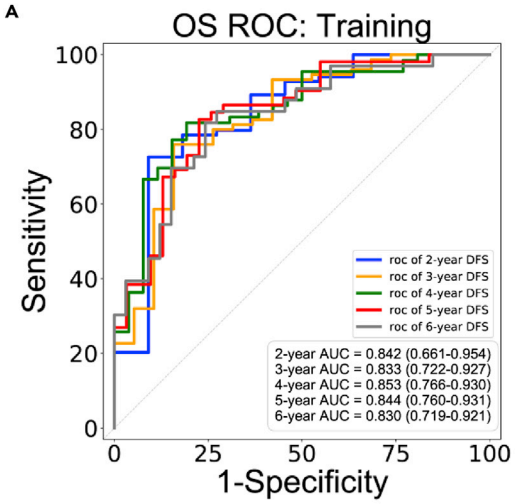
Figure 2. Frequency of radiomic features in FAMUS

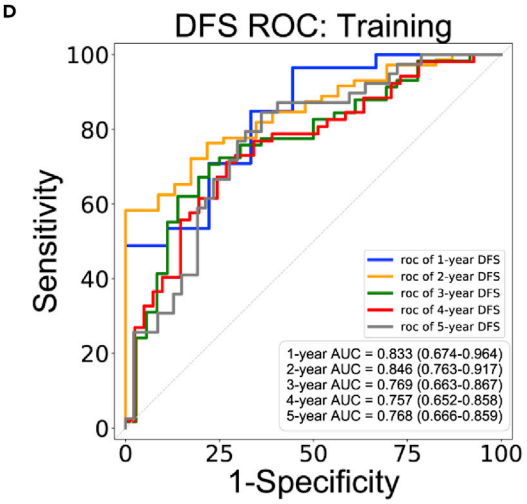
(A) OS-related features.

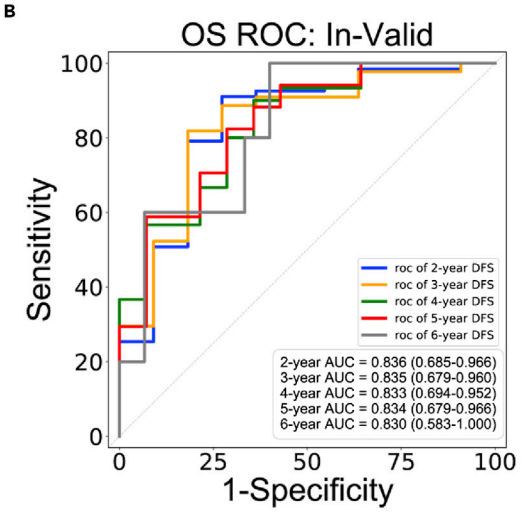
(B) DFS-related features.

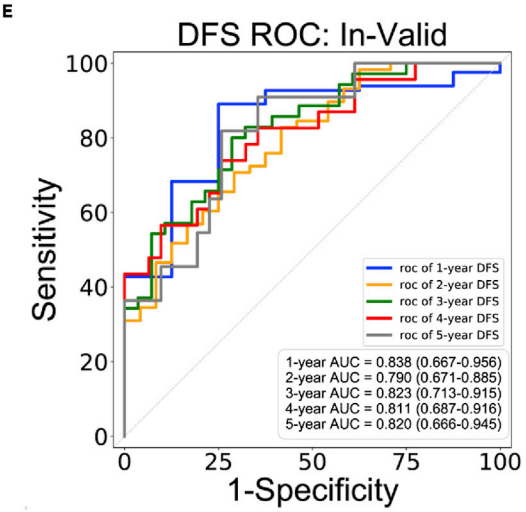
to be close to each other, but much higher than non-SASV features. Therefore, when sorting by the frequencies of these features, there should be some degree of sharp frequency change (SFC, similar to a change point) at the boundary between SASV and non-SASV features (Figure S3). Our Monte Carlo simulation analysis demonstrated that the first SFC can preserve sufficient features and thus can be used as a boundary to distinguish stable and unstable features (Table S2 and Figure S4). Using FAMUS, 13 and six SASV features were identified for OS and DFS, respectively. The frequency histograms indicated the appearance frequency of these features in FAMUS (Figures 2A and 2B).

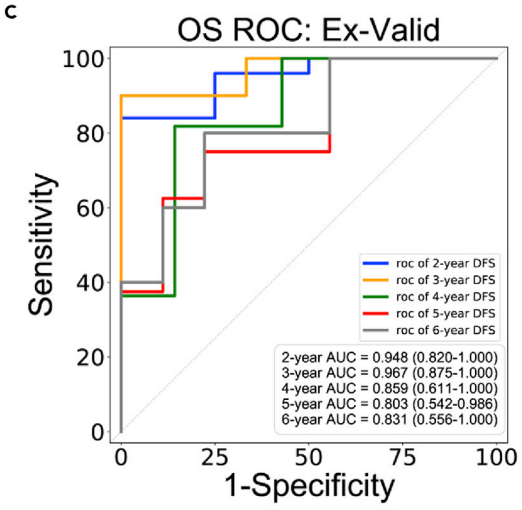
Thirdly, these features retained after the FAMUS procedure were further subjected to variable selection using the least absolute shrinkage and selection operator (LASSO) (Figure S5). As a result, two sets of three


A
OS ROC: Training
Sensitivity
1-Specificity
roc of 2-year DFS
roc of 3-year DFS
roc of 4-year DFS
roc of 5-year DFS
roc of 6-year DFS
2-year AUC = 0.842 (0.661-0.954)
3-year AUC = 0.833 (0.722-0.927)
4-year AUC = 0.853 (0.766-0.930)
5-year AUC = 0.844 (0.760-0.931)
6-year AUC = 0.830 (0.719-0.921)


D
DFS ROC: Training
Sensitivity
1-Specificity
roc of 1-year DFS
roc of 2-year DFS
roc of 3-year DFS
roc of 4-year DFS
roc of 5-year DFS
1-year AUC = 0.833 (0.674-0.964)
2-year AUC = 0.846 (0.763-0.917)
3-year AUC = 0.769 (0.663-0.867)
4-year AUC = 0.757 (0.652-0.858)
5-year AUC = 0.768 (0.666-0.859)


B
OS ROC: In-Valid
Sensitivity
1-Specificity
roc of 2-year DFS
roc of 3-year DFS
roc of 4-year DFS
roc of 5-year DFS
roc of 6-year DFS
2-year AUC = 0.836 (0.685-0.966)
3-year AUC = 0.835 (0.679-0.960)
4-year AUC = 0.833 (0.694-0.952)
5-year AUC = 0.834 (0.679-0.966)
6-year AUC = 0.830 (0.583-1.000)


E
DFS ROC: In-Valid
Sensitivity
1-Specificity
roc of 1-year DFS
roc of 2-year DFS
roc of 3-year DFS
roc of 4-year DFS
roc of 5-year DFS
1-year AUC = 0.838 (0.667-0.956)
2-year AUC = 0.790 (0.671-0.885)
3-year AUC = 0.823 (0.713-0.915)
4-year AUC = 0.811 (0.687-0.916)
5-year AUC = 0.820 (0.666-0.945)


C
OS ROC: Ex-Valid
Sensitivity
1-Specificity
roc of 2-year DFS
roc of 3-year DFS
roc of 4-year DFS
roc of 5-year DFS
roc of 6-year DFS
2-year AUC = 0.948 (0.820-1.000)
3-year AUC = 0.967 (0.875-1.000)
4-year AUC = 0.859 (0.611-1.000)
5-year AUC = 0.803 (0.542-0.986)
6-year AUC = 0.831 (0.556-1.000)

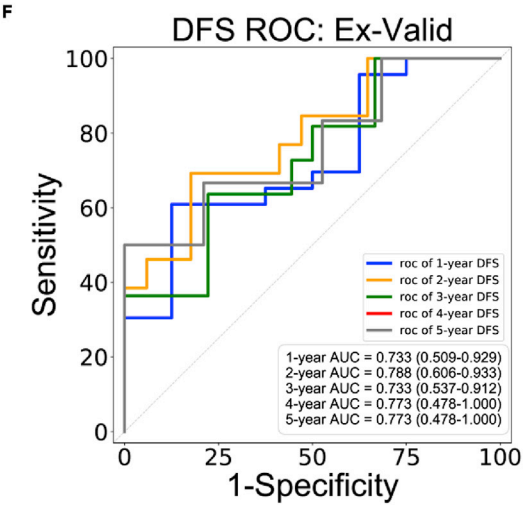

F
DFS ROC: Ex-Valid
Sensitivity
1-Specificity
roc of 1-year DFS
roc of 2-year DFS
roc of 3-year DFS
roc of 4-year DFS
roc of 5-year DFS
1-year AUC = 0.733 (0.509-0.929)
2-year AUC = 0.788 (0.606-0.933)
3-year AUC = 0.733 (0.537-0.912)
4-year AUC = 0.773 (0.478-1.000)
5-year AUC = 0.773 (0.478-1.000)

Figure 3. ROC curves of newly developed prognosis prediction models in three cohorts

(A–C) The performance of 2-6 years OS prediction in the training cohort (A), internal validation cohort (B), and external validation cohort (C).

(D–F) The performance of 1-5 years of DFS prediction in the training cohort (D), internal validation cohort (E), and external validation cohort (F).

different features were eventually selected from these stable features for OS and DFS, respectively (Table S3).

Construction of prognostic signatures

Next, using these stable and task-relevant features, radiomics-based signatures were identified to predict OS and DFS in patients with HGSOC using Cox proportional hazard models (Table S3). In the OS radiomic signatures, the hazard ratios of the three stable features were 1.35, 1.46, and 1.84, respectively. In the DFS radiomic signatures, the hazard ratios of the other three stable features were 1.34, 1.37, and 1.38, respectively. To faciliate query, the radiomic signatures of OS or DFS were computed by summing the raw and unnormalized scores of their respective three features multiplied with corresponding coefficients in the Cox models. The radiomic signatures could be formulated as:

$$\begin{aligned}\text{Radiomic signature for OS} = {} & 0.0127 \times \text{MNR-AL} \\ & + 0.0091 \times \text{HLH\_GLN1} \\ & + 3.0581 \times \text{LLL\_IMC2} \\ & - 3.7236\end{aligned}$$

$$\begin{aligned}\text{Radiomic signature for DFS} = {} & 86.9912 \times \text{HHL\_CT} \\ & + 0.4835 \times \text{SD} \\ & + 2.5690 \times \text{LLL\_IMC2} \\ & - 45.3635\end{aligned}$$

where the numerical coefficients were transformed by the inverse process of z-scores (Table S3).

Performance and validation of prognostic models

Then, the predictive performance of the radiomic models was assessed using the concordance index (C-Index) in the training set, internal validation set, and external validation set. For OS, the C-Index was 0.791 (95% confidence interval (CI:) 0.706–0.857), 0.816 (95% CI: 0.674–0.924) and 0.858 (95% CI: 0.707–0.942) in the training, internal validation, and external validation cohorts, respectively. For DFS, the C-Index was 0.734 (95% CI: 0.662–0.794), 0.754 (95% CI: 0.666–0.833) and 0.700 (95% CI: 0.569–0.813) in the same three cohorts, respectively. The receiver operation characteristics (ROC) curves showing the performance of radiomics-based classifier at various classification thresholds were plotted for 2–6 years OS and 1–5 years DFS in the three cohorts (Figure 3). The area under the ROC curves (AUC) for predicting 5-year OS was 0.844 (95% CI: 0.760–0.931) in the training cohort, 0.834 (95% CI: 0.679–0.966), and 0.803 (95% CI: 0.542–0.986) in the internal and external validation cohort, respectively (Figures 3A–3C). The AUC for predicting the 3-year DFS was 0.769 (95% CI: 0.663–0.867) in the training cohort, 0.823 (95% CI: 0.713–0.915), and 0.733 (95% CI: 0.537–0.912) in the internal and external validation cohorts, respectively (Figures 3D–3F). Overall, all the median AUCs for OS were higher than 0.80 and all median AUCs for DFS were higher than 0.73. The high C-indexes and AUCs for both OS and DFS in all cohorts indicated the good prognostic performance of the developed radiomics-based prediction models.

Kaplan-Meier survival curves were also plotted for groups stratified by the risk of 5-year OS or 3-year DFS predicted by their radiomic models (Figure 4). The predicted low-risk group of patients had a significantly longer OS or DFS than the high-risk group in both internal and external validation cohorts (p value <0.01, Log rank test). Similarly, other Kaplan-Meier curves of high- and low-risk groups, stratified by different survival predictions using the developed radiomic models (Figures S6–S9), also showed significant differences in OS or DFS between their two corresponding stratified groups in both of the validation cohorts (all p values <0.05, Log rank test). These results consistently demonstrated that the developed models were robust and of clinical utility in stratifying high-risk patients for the adjustment of treatment strategies according to their preoperative CT images.

The best cut-off values, as determined by the Youden Index for low- and high-risk group stratification, are listed in Table S4. Sensitivity, specificity, PPV, and NPV for 5-year OS and 3-year DFS under their corresponding cut-off values are listed in Table 2. For example, for 5-year OS in which patients were predicted to live

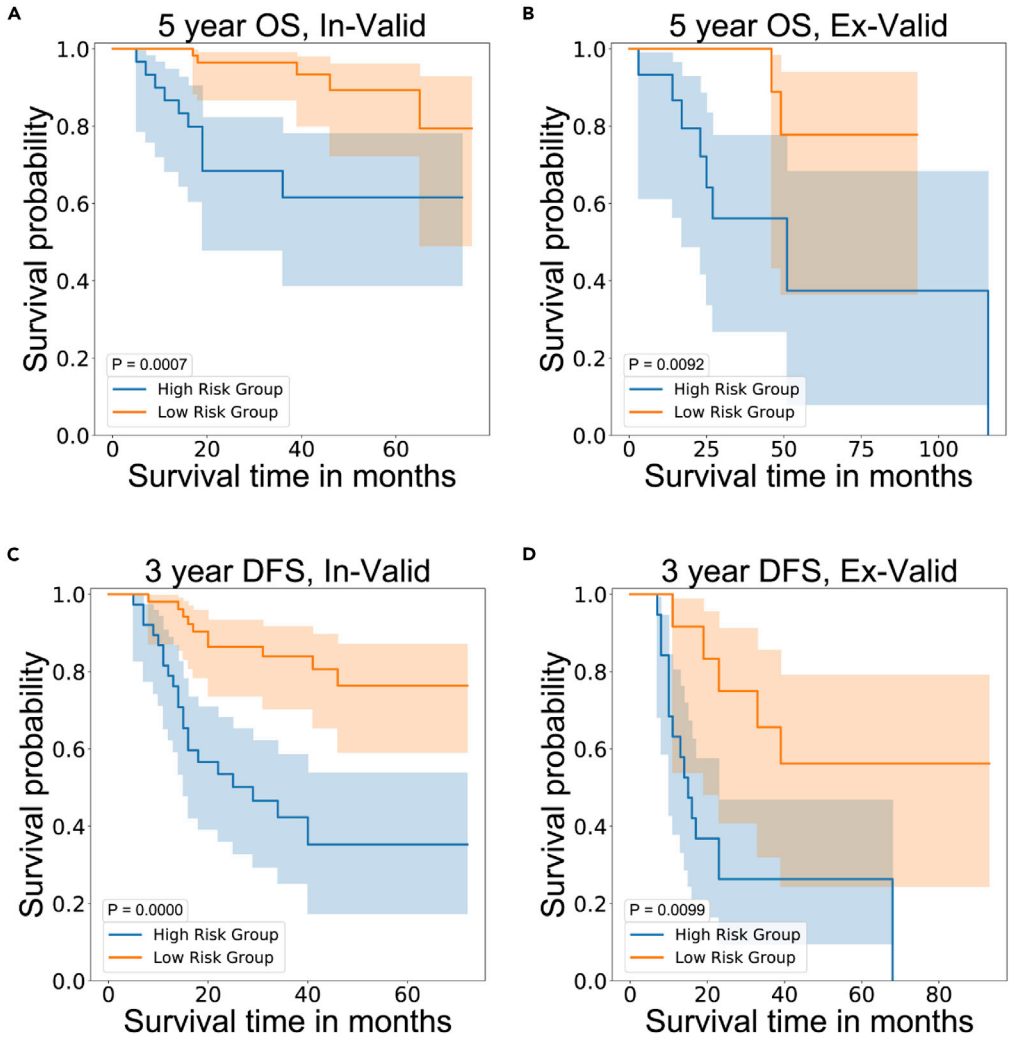


Figure 4. Kaplan-Meier curves of high-risk and low-risk groups stratified by the developed radiomics prediction models

(A and B) 5-year OS prediction in the internal validation cohort (A) and external validation cohort (B). For 5-year OS prediction, patients that were predicted to live longer than 5-year OS were assigned as the low-risk group, otherwise they were assigned as the high-risk group.

(C and D) 3-year DFS prediction in the internal validation cohort (C) and external validation cohort (D). For 3-year DFS prediction, patients that were predicted to live longer than 3 years with DFS were assigned as the low-risk group, otherwise they were assigned as the high-risk group. Log rank tests were used to compare the differences between the survival curves of these paired groups.

longer than 5-year OS, these were assigned to the low-risk group, and the rest of the patients were assigned to the high-risk group. For this, the sensitivity, specificity, PPV, and NPV were 0.750, 0.778, 0.750, and 0.778 in the external validation cohort, respectively. For 3-year DFS, patients that were predicted with longer than 3-year DFS were assigned into the low-risk group and otherwise assigned into the high-risk group. Here the sensitivity, specificity, PPV, and NPV were 0.636, 0.778, 0.636, and 0.778 in the external validation cohort, respectively. These metrics also proved the good performance of these developed radiomic modes.

Radiomic nomograms

Nomogram models that incorporate the corresponding radiomic signatures for predicting OS and DFS were established (Figures 5A and 5C). In these nomogram models, the probability of 2–6-year survival for OS or 1–5-year survival for DFS could be queried. It is notable that the coefficient values of the radiomic

Table 2. Sensitivity, specificity, PPV, and NPV for 5-year OS and 3-year DFS under cut-off value determined using the Youden Index

	Cut-off value	Sensitivity	Specificity	PPV[a]	NPV
5-year OS					
Training cohort	0.699	0.827	0.774	0.860	0.727
Internal validation cohort	0.776	0.824	0.714	0.778	0.769
External validation cohort	0.710	0.750	0.778	0.750	0.778
3-year DFS					
Training Cohort	0.692	0.707	0.778	0.837	0.622
Internal validation cohort	0.795	0.800	0.714	0.778	0.741
External validation cohort	0.612	0.636	0.778	0.636	0.778

[a]PPV: positive predictive value; NPV: negative predictive value.

signature formulas were quite different because all the feature values were original and unnormalized values. This is to facilitate the querier to directly use the original feature values instead of requiring them to calculate the normalized values. In addition, the parameters of each feature for the *Z* score are listed in Table S3.

Furthermore, the calibration curve of the radiomic nomogram used to estimate OS or DFS outcomes were tightly distributed along the diagonal (Figures 5B and 5D), implying good agreement between prediction and observation in training, internal, and external validation sets.

Analysis of features with the stability against sample variation

Finally, the impact of FAMUS on the performance of the developed radiomic models was assessed. Four different feature selection strategies with FAMUS, that is, (i) SCC + FAMUS + LASSO, (ii) SCC + FAMUS + SFFS, (iii) SCC + FAMUS + ReliefF, and (iv) SCC + FAMUS + mRMR, were compared with their counterparts without FAMUS, that is, (i) SCC + LASSO, (ii) SCC + SFFS, (iii) SCC + ReliefF, and iv) SCC + mRMR. In all feature selection strategies, redundant features were first filtered by SCC. Then, in the four strategies with FAMUS, the unique features obtained via the SCC filtering were further selected by our newly proposed FAMUS before the processing of LASSO (i.e., SCC + FAMUS + LASSO), SFFS (i.e., SCC + FAMUS + SFFS), ReliefF (i.e., SCC + FAMUS + ReliefF), or mRMR (i.e., SCC + FAMUS + mRMR). In the four strategies without FAMUS, the unique features obtained after the SCC filtering were directly processed by LASSO (i.e., SCC + LASSO), SFFS (i.e., SCC + SFFS), ReliefF (i.e., SCC + ReliefF), or mRMR (i.e., SCC + mRMR).

At the step of FAMUS, the median frequency of features for OS and DFS was 29.1% (IQR 20.1–44.0%) and 35.3% (IQR 26.5–56.9%), respectively. Only 17% of features for OS and 8% of features for DFS were retained in the study, indicating that most features were unstable to sample variation (Figure 2).

Average Hamming Distance (AHD) was calculated to measure the stability of the feature selection results. Table 3 lists the AHD obtained from different methods. The four methods with FAMUS had much smaller AHDs than their counterparts without FAMUS, suggesting that FAMUS could greatly improve the stability of the feature selection algorithm.

Incorporating these stable features into the radiomic models improved the efficiency of survival prediction in several aspects. Firstly, for both OS and DFS, the mean C-indexes of the four FAMUS-containing methods were significantly higher than those of their corresponding non-FAMUS-containing methods in both internal and external validation datasets (all p values < 0.001, paired t-test) (Figure 6). Secondly, compared with SCC + LASSO, the performance of the other non-FAMUS-containing methods was significantly worse, but the performance of adding FUMUS was improved to be close or even better than SCC + FAMUS + LASSO. It was also indicated that adding FAMUS to Filter (e.g., ReliefF and mRMR) or Wrapper (e.g., SFFS) methods could benefit more than Embedded (e.g., LASSO) methods. Thirdly, the box bodies of non-FAMUS-containing methods were longer than those of FAMUS-containing methods in all datasets for all tasks, indicating that the latter four methods were more robust than the former four methods.

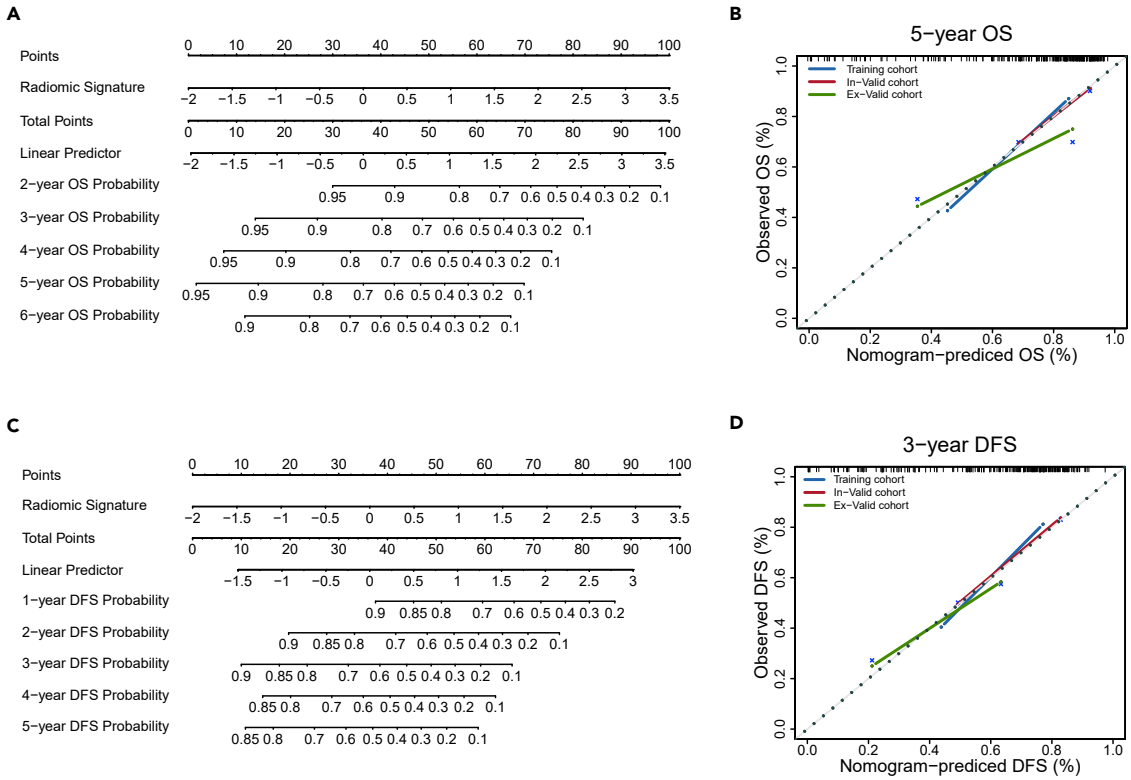


Figure 5. Radiomics nomograms constructed by the prognosis prediction models

(A) Radiomics nomogram to estimate the risk of cancer death in patients with HGSOC.

(B) 5-year OS calibration curves of the radiomics nomogram in the training cohort and combined validation cohort.

(C) Radiomics nomogram to estimate the risk of cancer recurrence in patients with HGSOC.

(D) 3-year DFS calibration curves of the radiomics nomogram in the training cohort and combined validation cohort.

Furthermore, the mean C-indexes of the four non-FAMUS-containing methods between the training and validation datasets were less consistent than those of their corresponding FAMUS-containing methods, indicating that adding FAMUS to the construction of prognostic signatures could reduce model overfitting. All these results demonstrated that FAMUS could substantially improve the performance of the survival prediction model based on feature selection methods.

DISCUSSION

In this study, the prediction models of OS and DFS in patients with HGSOC were developed based on SASV radiomic features. The effectiveness of the two models was evaluated using independent internal and external validation datasets. Notably, as stability of the selected features against sample variation was considered, it was demonstrated that the addition of SASV features resulted in a significant improvement in the performance of the models. This could help to guide future clinical decision-making processes in a reliable and reproducible fashion. Specifically, by using the developed prediction models, patients with HGSOC can be accurately stratified into high-risk or low-risk groups of 2–6 years predictions for cancer-related death or 1–5 years for likely recurrence, according to their CT images. This can give more detailed information for clinicians to formulate treatment plans, providing more aggressive treatment for high-risk patients and less aggressive treatment for low-risk patients. Finally, two nomograms based on the two corresponding radiomic signatures were developed, offering a low-cost, non-invasive means for predicting the risk for HGSOC cancer-related death or relapse.

Studies have shown that radiomic analysis is feasible as a non-invasive prediction tool of prognosis based on CT images (Rathore et al., 2018; Wu et al., 2017; Xie et al., 2019). Some previous radiomic analyses of ovarian cancer extracted features only from primary lesions (Lu et al., 2019; Wei et al.,

Table 3. AHDs of eight different methods for building OS and DFS prognosis prediction models

Methods	AHD for OS	AHD for DFS
SCC + FAMUS + LASSO	3.533	2.258
SCC + FAMUS + SFFS	5.021	2.351
SCC + FAMUS + ReliefF	2.713	2.035
SCC + FAMUS + mRMR	3.155	1.137
SCC + LASSO	8.460	8.456
SCC + SFFS	12.688	12.845
SCC + ReliefF	4.667	3.811
SCC + mRMR	6.959	7.482

2019). However, metastatic lesions have also been significantly correlated with both OS and DFS (Leffers et al., 2009). In our cohorts, patients with metastatic lesions had dramatically different survival curves from ones without metastatic lesions (Figure S10). Therefore, metastatic lesions were included in our subregions for feature extraction. Totally, 891 radiomic features were extracted from CT images in the study, and SCC was then used to exclude redundant features which may slow down the training process and adversely affect the performance of the prediction models. After the SCC analysis, 130 and 93 features remained for OS and DFS, respectively. In other words, about 85% (761/891) of the features for OS and 89.6% (798/891) of the features of DFS were regarded as redundant. During feature selection, the stability of features against sample variation was further considered, and a heuristic method, termed FAMUS, was proposed to determine stable features. As a result, 12 features of OS and six features of DFS were judged by FAMUS to have particularly high stability in sample variation. Finally, LASSO, a high-performance feature selection algorithm commonly used in machine learning, was used to select features closely related to the risk of cancer death or recurrence. Two different sets of three features were determined to establish accurate prediction models for OS and DFS, respectively.

Both OS and DFS prediction models showed high prognostic values with a C-index of 0.791–0.858 and 0.700–0.754 in the training, internal, and external validation cohorts, respectively. The consistent results achieved in the validation cohorts indicated the stability of both OS and DFS models in independent datasets. Both 2–6 years OS ROC curves and 1–5 years DFS ROC curves (Figures 3A–3F) visualized the high performance of the two models. Furthermore, high-risk and low-risk groups of 2-6 years cancer-related death and 1–5 years relapse were successfully stratified according to the prediction results. An OS radiomic signature and a DFS radiomic signature were built using coefficients of features in the prediction models. The two radiomic signatures were then used to develop a radiomic nomogram of OS and DFS, respectively.

Though these newly developed models can provide clinicians with strong prognostic information before therapy and help tailor treatment strategies for patients, it was worth noting that the performance of radiomic analysis in DFS was worse than that in OS, which is in good agreement with previous studies (Lu et al., 2019; Wei et al., 2019). This is perhaps owing to the small number and low frequency of the SASV features of DFS in FAMUS compared with OS (Figure 2), which highlighted the weakness of using only radiomic analysis in DFS. Further studies are needed to combine more clinical data with radiomic features for DFS to develop better radiomic prediction models of DFS.

Previous studies on feature stability have mainly focused on the effect of different scanning settings on radiomics features (Wei et al., 2019; Wu et al., 2017). One study focused on the influence of radiomics itself on reliable prognostic results. Though the effect of dataset changes on feature selection results has been studied to some extent (Pes, 2017), they have yet to be applied to practical radiomic analysis. FAMUS proposed in the study is an effective algorithm to find prognostic features with stability against sample variation. For our specific tasks in prognosis, the univariate Cox regression was taken as the prescreening method in the FAMUS step. The stable features selected by FAMUS tended to have a stronger correlation with prognosis, whereas less stable features lacked this correlation. As shown in Figures 6A and 6B, the models trained on stable features had better performance than those trained

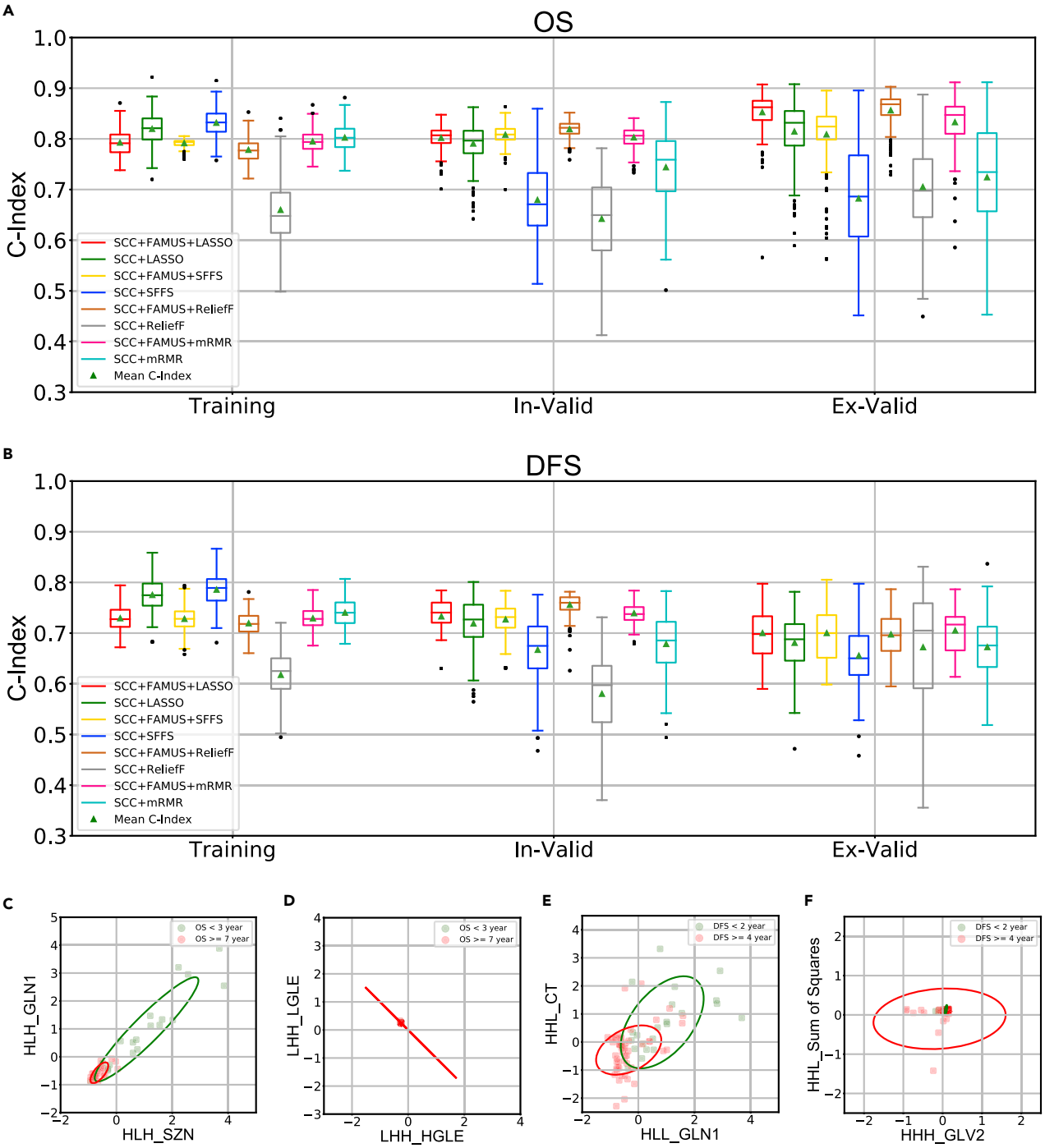


Figure 6. Impacts of FAMUS on the performance of prognosis prediction models

Eight different methods used for constructing the prognosis prediction models are compared, including SCC + FAMUS + LASSO, SCC + LASSO, SCC + FAMUS + SFFS, SCC + SFFS, SCC + FAMUS + ReliefF, SCC + ReliefF, SCC + FAMUS + mRMR and SCC + mRMR.

(A) OS radiomics prediction model. (B) DFS radiomics prediction model. Data are shown in boxplots. The thick line in the box is median and box spans from Q1 (25th percentile) to Q3 (75th percentile). The whiskers extend to the most extreme observation within 1.5 times the interquartile range (Q3–Q1) from the nearest quartile.

(C–F) Scatterplots displaying the relationship between two stable features of OS (C) or DFS (E), and the relationship between two unstable features of OS (D) or DFS (F).

on features selected directly from SCC. Interestingly, stable features showed distinct patterns between high-risk and low-risk groups stratified by either OS or DFS (Figures 6C and 6E), while unstable features did not have such distinct patterns between high-risk and low-risk groups of survival (Figures 6D and 6F).

In conclusion, the newly developed prognosis models and their associated nomograms based on radiomics features have great performance in predicting the risk of OS or DFS for patients with HGSOC treated with primary debulking surgery followed by neoadjuvant chemotherapy. The improved performance of these prognosis prediction models is mainly attributed to the heuristic FAMUS that was used for selecting radiomics features with the stability against sample variation. It is proposed as a convincing and reliable model for use in radiomic signature analysis based on stable features as a non-invasive prognostic marker for the individualized evaluation of patients with HGSOC.

Limitations of the study

Several study caveats should be acknowledged. Firstly, this is a retrospective study, and the validation cohorts are of relatively small size. The proportion of cancer-related death cases in the internal validation cohorts is also small. Secondly, other types of images like magnetic resonance spectroscopy and other relevant pathological information normally used in clinical practice were not included in the study. Prospective trials are thus required in future studies to address these limitations. Thirdly, the cut-off value of FAMUS was determined empirically and subjectively, which means it may not be optimal and some stable features may have been removed. Although stable features selected by FAMUS exhibited excellent performance in the prognostic prediction in the present study, an objective method to determine the exact cut-off value should be fine-tuned in further studies to solve these limitations. Lastly, in a study, based upon a fairly localized Chinese cohort, any potential racial or geographic disparities in histological and clinical features may also require consideration before global application.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - Lead contact
  - Materials availability
  - Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Patient recruitment
  - Follow-up times
- METHOD DETAILS
  - Tumor segmentation and pre-processing
  - Feature extraction
  - Feature selection by SCC
  - Feature selection by FAMUS
  - Feature selection by LASSO
  - Construction of radiomic signature
  - Prognostic performance evaluation
  - Comparison of the feature selection methods
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2022.104628.

ACKNOWLEDGMENTS

We thank the Core Facility at Zhejiang University School of Medicine for providing technical support. We also thank Dr. Christopher R. Wood for English grammar polishing. This work has been supported in part by the Key R&D Program of Zhejiang Province (2021C03126), Medical Health Science and Technology Key Project of Zhejiang Provincial Health Commission (WKJ-ZJ-2007), the Key Program of Zhejiang Provincial

Natural Science Foundation (LZ20H160001), National Natural Science Foundation of China (82072857), and the Leading Innovative and Entrepreneur Team Introduction Program of Zhejiang of China (2019R01001).

AUTHOR CONTRIBUTIONS

Y.L. and P.L. considered and designed the study. J.H. performed primary data analyses. R.Z. and C.Z. performed secondary data analysis. Z.W., B.L., and W.L. performed tumor labeling on CT images. Y.L. and C.Z. collected CT image data. X.C. collected clinical data. J.H. wrote the article. P.L. and Y.L. revised the article. All of the authors discussed and commented on the study.

DECLARATION OF INTERESTS

The authors declare no competing interests.

INCLUSION AND DIVERSITY

We worked to ensure ethnic or other types of diversity in the recruitment of human subjects. While citing references scientifically relevant for this work, we also actively worked to promote gender balance in our reference list.

Received: December 22, 2021
Revised: May 3, 2022
Accepted: June 13, 2022
Published: July 15, 2022

REFERENCES

Aerts, H.J.W.L., Velazquez, E.R., Leijenaar, R.T.H., Parmar, C., Grossmann, P., Cavalho, S., Bussink, J., Monshouwer, R., Haibe-Kains, B., Rietveld, D., et al. (2014). Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. Nat. Commun. *5*, 5006.

Bowtell, D.D., Bohm, S., Ahmed, A.A., Aspuria, P.J., Bast, R.C., Jr., Beral, V., Berek, J.S., Birrer, M.J., Blagden, S., Bookman, M.A., et al. (2015). Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. Nat. Rev. Cancer *15*, 668–679.

Chen, W., Zheng, R., Baade, P.D., Zhang, S., Zeng, H., Bray, F., Jemal, A., Yu, X.Q., and He, J. (2016). Cancer statistics in China, 2015. CA A Cancer J. Clin. *66*, 115–132.

Dunne, K., Cunningham, P., and Azuaje, F. (2002). Solutions to instability problems with sequential wrapper-based approaches to feature selection. J. Mach. Learn. Res. *1*, 1–22.

Grant, S.W., Hickey, G.L., and Head, S.J. (2019). Statistical primer: multivariable regression considerations and pitfalls. Eur. J. Cardio Thorac. *55*, 179–185.

Harrell, F.E., Lee, K.L., and Mark, D.B. (1996). Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. Stat. Med. *15*, 361–387.

Hu, T.D., Wang, S.P., Huang, L., Wang, J.Z., Shi, D.B., Li, Y., Tong, T., and Peng, W.J. (2019). A clinical-radiomics nomogram for the preoperative prediction of lung metastasis in colorectal cancer patients with indeterminate pulmonary nodules. Eur. Radiol. *29*, 439–449.

Kalousis, A., Prados, J., and Hilario, M. (2007). Stability of feature selection algorithms: a study on high-dimensional spaces. Knowl. Inf. Syst. *12*, 95–116.

Leffers, N., Gooden, M.J., de Jong, R.A., Hoogeboom, B.N., ten Hoor, K.A., Hollema, H., Boezen, H.M., van der Zee, A.G., Daemen, T., and Nijman, H.W. (2009). Prognostic significance of tumor-infiltrating T-lymphocytes in primary and metastatic lesions of advanced stage ovarian cancer. Cancer Immunol. Immunother. *58*, 449–459.

Lu, H.N., Arshad, M., Thornton, A., Avesani, G., Cunnea, P., Curry, E., Kanavati, F., Liang, J., Nixon, K., Williams, S.T., et al. (2019). A mathematical-descriptor of tumor-mesoscopic-structure from computed-tomography images annotates prognostic- and molecular-phenotypes of epithelial ovarian cancer. Nat. Commun. *10*, 764.

Nougaret, S., Lakhman, Y., Gonen, M., Goldman, D.A., Micco, M., D'Anastasi, M., Johnson, S.A., Juluru, K., Arnold, A.G., Sosa, R.E., et al. (2017). High-grade serous ovarian cancer: associations between BRCA mutation status, CT imaging phenotypes, and clinical outcomes. Radiology *285*, 472–481.

Pes, B. (2017). Feature selection for high-dimensional data: the issue of stability. In 26th International Conference on Enabling Technologies: Infrastructure for Collaborative Enterprises (WETICE) (IEEE), pp. 170–175.

Pupo, O.G.R., Morell, C., and Soto, S.V. (2013). In ReliefF-ML: An Extension of ReliefF Algorithm to Multi-label Learning, J. Ruiz-Shulcloper and G. Sanniti di Baja, eds. (Springer Berlin Heidelberg), pp. 528–535.

Rathore, S., Akbari, H., Rozycki, M., Abdullah, K.G., Nasrallah, M.P., Binder, Z.A., Davuluri, R.V., Lustig, R.A., Dahmane, N., Bilello, M., et al. (2018). Radiomic MRI signature reveals three distinct subtypes of glioblastoma with different clinical and molecular characteristics, offering prognostic value beyond IDH1. Sci. Rep. *8*, 5087.

Sung, H., Ferlay, J., Siegel, R.L., Laversanne, M., Soerjomataram, I., Jemal, A., and Bray, F. (2021). Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *71*, 209–249.

Tan, M.X., Pu, J.T., and Zheng, B. (2014). Optimization of breast mass classification using sequential forward floating selection (SFFS) and a support vector machine (SVM) model. Int. J. Comput. Assist. Rad. *9*, 1005–1020.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. Stat. Med. *16*, 385–395.

Vallieres, M., Freeman, C.R., Skamene, S.R., and El Naqa, I. (2015). A radiomics model from joint FDG-PET and MRI texture features for the prediction of lung metastases in soft-tissue sarcomas of the extremities. Phys. Med. Biol. *60*, 5471–5496.

van Griethuysen, J.J.M., Fedorov, A., Parmar, C., Hosny, A., Aucoin, N., Narayan, V., Beets-Tan, R.G.H., Fillion-Robin, J.C., Pieper, S., and Aerts, H.J.W.L. (2017). Computational radiomics system to decode the radiographic phenotype. Cancer Res. *77*, E104–E107.

van Timmeren, J.E., Leijenaar, R.T.H., van Elmpt, W., Reymen, B., Oberije, C., Monshouwer, R., Bussink, J., Brink, C., Hansen, O., and Lambin, P. (2017). Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiother. Oncol. *123*, 363–369.

Vang, R., Shih Ie, M., and Kurman, R.J. (2009). Ovarian low-grade and high-grade serous carcinoma: pathogenesis, clinicopathologic and molecular biologic features, and diagnostic problems. Adv. Anat. Pathol. *16*, 267–282.

Vaughan, S., Coward, J.I., Bast, R.C., Jr., Berchuck, A., Berek, J.S., Brenton, J.D., Coukos, G., Crum, C.C., Drapkin, R., Etemadmoghadam, D., et al. (2011). Rethinking ovarian cancer: recommendations for improving outcomes. Nat. Rev. Cancer *11*, 719–725.

Wei, W., Liu, Z.Y., Rong, Y., Zhou, B., Bei, Y., Wei, W., Wang, S., Wang, M.Y., Guo, Y.K., and Tian, J. (2019). A computed tomography-based radiomic prognostic marker of advanced high-grade serous ovarian cancer recurrence: a multicenter study. Front. Oncol. *9*, 255.

Wu, S.X., Zheng, J.J., Li, Y., Yu, H., Shi, S.Y., Xie, W.B., Liu, H., Su, Y.F., Huang, J., and Lin, T.X. (2017). A radiomics nomogram for the preoperative prediction of lymph node metastasis in bladder cancer. Clin. Cancer Res. *23*, 6904–6911.

Hao, X., Liu, B., Hu, X., Wei, J., Han, Y., Liu, X., Chen, Z., Li, J., Bai, J., Chen, Y., et al. (2021). A radiomics-based approach for predicting early recurrence in intrahepatic cholangiocarcinoma after surgical resection: a multicenter study. Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. *2021*, 3659–3662.

Xie, C.Y., Yang, P.F., Zhang, X.B., Xu, L., Wang, X.J., Li, X.D., Zhang, L.H., Xie, R.F., Yang, L., Jing, Z., et al. (2019). Sub-region based radiomics analysis for survival prediction in oesophageal tumours treated by definitive concurrent chemoradiotherapy. EBioMedicine *44*, 289–297.

Yu, L., and Liu, H. (2003). Feature selection for high-dimensional data: a fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), pp. 856–863.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
ITK-SNAP	ITK-SNAP	http://www.itksnap.org/pmwiki/pmwiki.php?n=Downloads.SNAP3
Radiomic feature extraction	pyradiomics	https://pyradiomics.readthedocs.io/en/latest/
LASSO	scikit-survival	https://pypi.org/project/scikit-survival/
SFFS	github	https://github.com/AKittenOfMrHu/SFFS_for_survival
ReliefF	ReliefF	https://pypi.org/project/ReliefF/
mRMR	mrmr-selection	https://pypi.org/project/mrmr-selection/
Survival prediction model	lifelines	https://lifelines.readthedocs.io/en/latest/
Statistical analysis by Python	scipy	https://scipy.org/
Statistical analysis by R	R	http://www.R-project.org
nomogram model	R	https://cran.r-project.org/web/packages/rms/ https://cran.r-project.org/web/packages/survival/index.html

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Yan Lu (yanlu76@zju.edu.cn).

Materials availability

This study did not generate new materials.

Data and code availability

- All data generated during this study are included in this published article and its supplemental information.
- This paper does not report original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Patient recruitment

A total of 217 patients diagnosed with HGSOC were enrolled in this study. Patient inclusion criteria were 1) pathologically confirmed primary HGSOC, 2) both OS and DFS data fully available, and 3) preoperative venous phase contrast enhanced CT of the abdomen and pelvis conducted and available. Patient exclusion criteria were 1) undergoing neoadjuvant chemotherapy before primary debulking surgery, 2) non-treatment related death, 3) no qualified CT within two months before surgery, or 4) CT with high noise or artifacts in the lesion. Patients from the Women's Hospital of Zhejiang University School of Medicine were divided into a training cohort comprised of 95 patients (diagnosed between January 2008 and November 2012), and an internal validation cohort comprised of 90 patients (diagnosed between January 2013 and November 2018). Thirty-two patients (diagnosed between Match, 2009 and November 2017 at the First Affiliated Hospital of Wenzhou Medical University) were used as an external independent validation cohort. The median age of patients in the training, internal validation, and external validation cohorts was 50 (20–73), 51.5 (18–73), and 55 (32, 68), respectively (Table 1). There was no significant age difference between the three cohorts (p-value = 0.450). This study was reviewed and approved by the Institutional Review Board of both hospitals. The study was conducted in accordance with the International Ethical Guidelines for Biomedical Research Involving Human Subjects.

Follow-up times

The outcome variables are OS (the time from the date of primary debulking surgery to the date of death) and DFS (the time from the date of primary debulking surgery to the first evidence of locoregional or distant recurrence). All patients were followed up yearly and were regularly evaluated at the end of their treatment for evidence of disease recurrence. Recurrence dates were determined according to a follow-up physical exam, CT findings, and CA-125 levels.

METHOD DETAILS

Tumor segmentation and pre-processing

Regions of interest (ROI) were segmented by three radiologists with 17, 11, and 17 years' experience in interpreting CT images of ovarian cancer separately using the ITK-SNAP software (http://www.itksnap.org/pmwiki/pmwiki.php), respectively. Feature extraction was then performed based on three segmented tumor subregions, namely primary, metastatic, and lymphatic lesions. The slice in-plane spacing ranged from 1.25 mm to 10.00 mm. To compensate for differences in radiomic features caused by different reconstruction slice thicknesses and pixel sizes, the voxel sizes of all CT images in this study were reconstructed to $1 \times 1 \times 10\ mm^3$.

Feature extraction

Before the feature extraction, CT intensity values were limited to between −105 and 195 Hounsfield Units (window level: 45, window width: 300) and were then normalized to 64 gray levels by linear mapping. A total of 851 radiomic features in 7 categories were extracted for each patient based on the patient's ROI using an open-source Python package "pyradiomics" (van Griethuysen et al., 2017). Details on these features can be found in Table S1 and Figure S2 with more comprehensive definitions and descriptions available in the study of Vallieres et al. (Vallieres et al., 2015).

Feature selection by SCC

Feature selection is a key step for high-dimensional data analysis in radiomics. It is a method of making the overall analysis more manageable, efficient, and productive by eliminating irrelevant and 'noisy' features. The Z-score was used to perform data normalization before feature selection. The Spearman correlation coefficient (SCC) was first taken to determine redundant features. When SCC >0.8 with p value between two features <0.05, the one with larger p value calculated by univariate Cox regression was regarded as a redundant and removed.

Feature selection by FAMUS

The stability of the features after the SCC filtering to sample variation was evaluated. This is a very important issue and seeking a solution should be a high priority. However, it seems this issue has been rarely considered in previous studies. Even if the predictive performance is good in both training and validation datasets, this contingency cannot be ignored. A feature is regarded as stable if it is always selected when samples in the dataset varies. To this end, a heuristic method, FAMUS, was developed to identify features with good stability against sample variation (Figure 1). To obtain more general subset of data, the training data was sampled from failing and censored patients, respectively. At the sampling rate ρ and sampling time N, the frequency of each feature appearing in the selection process was calculated.

Suppose that the probability distribution of optimal selected features $\boldsymbol{f}$ from data space $\boldsymbol{D}$ is $p(\boldsymbol{f})$, and the probability distribution of selected features $\widehat{\boldsymbol{f}}$ from experimental data $\widehat{\boldsymbol{D}}$ is $p(\widehat{\boldsymbol{f}}|\widehat{\boldsymbol{D}})$. Therefore, differently sampled datasets will result in different selected features. Ideally, $\boldsymbol{f} = \widehat{\boldsymbol{f}}$ and $p(\boldsymbol{f}) = \int p(\widehat{\boldsymbol{f}}|\widehat{\boldsymbol{D}})p(\widehat{\boldsymbol{D}})\mathrm{d}\widehat{\boldsymbol{D}}$. Due to the limitation of experimental data, FAMUS aims to calculate approximate distribution $g(\widehat{\boldsymbol{f}})$ asymptotically equal to $p(\boldsymbol{f})$. Monte Carlo simulation was employed to differently sampled datasets $\widehat{\boldsymbol{D}}_i$, $i \in \{1, 2, \ldots, N\}$ from $\widehat{\boldsymbol{D}}$ with a sampling rate ρ. At the sampling rate ρ and sampling time N, $g(\widehat{\boldsymbol{f}}) = \sum_{i=1}^{N} p(\widehat{\boldsymbol{f}}|\widehat{\boldsymbol{D}}_i)p(\widehat{\boldsymbol{D}}_i)$.

Two hyper-parameters, the sampling rate ρ and sampling time N, are related the above equation $g(\widehat{\boldsymbol{f}})$. There should be sufficient variation between differently sampled datasets. As an extreme example, $\widehat{\boldsymbol{D}}_i = \widehat{\boldsymbol{D}}$ is equivalent to one time of univariable prescreening. In other words, ρ should be small to reflect the differences between stable and task-relevant features and other features. To determine the

appropriate sampling rate ρ, in our Monte Carlo simulations, N was set to 2000 and ρ to 1/6, 1/5, 1/4 or 1/3. The sharper frequency changes were observed in the experiments with ρ of 1/6 and 1/5 comparing to those with ρ of 1/4 and 1/3 (Figure S3). Therefore, ρ of 1/6 was chosen in the study.

The sampling time N should be large so that $\boldsymbol{g}(\widehat{\boldsymbol{f}})$ approximates to $\boldsymbol{p}(\boldsymbol{f})$. When ρ was set to 1/6, simulation experiments of N of 800, 900, 1000, 1500 and 2000 were implemented. According to their frequencies, the selected 13 OS features or 6 DFS features in the study were always the top 13 or the top 6 features in these experiments with N of 2000. Furthermore, similar results were observed in experiments with ρ of 1/6, 1/5, 1/4 or 1/3 and with N of 2000. Therefore, N was set to 2000 in the study.

In addition, the univariable prescreening included only covariates that were significant at a particular threshold based on a univariable model. However, a commonly used threshold is p value <0.05, which may result in the removal of important covariate variables from the model due to stochastic variability. It was reflected that when ρ was 1/6 and N was 2000, the frequencies of features with p value <0.05 in FAMUS were much lower than those of features with p value <0.25. The median (the interquartile ranges, IQR) frequencies of the top 10 features with p value <0.05 in FAMUS were 0.367 (0.331, 0.495) for OS and 0.312 (0.285, 0.361) for DFS, suggesting that most features had frequencies lower than 0.5. The median (IQR) frequencies of the top 10 features with p value <0.2 in FAMUS were 0.761 (0.743, 0.832) for OS and 0.709 (0.668, 0.735) for DFS. Other researchers empirically recommended p values of 0.20 and 0.25 for feature selections (Grant et al., 2019). Therefore, a p value <0.25 instead of <0.05 was taken as threshold for FAMUS prescreening to avoid removing important covariates from the model due to random variability.

Feature selection by LASSO

The FAMUS-filtered features were selected using the least absolute shrinkage and selection operator (LASSO) for Cox regression analysis (Tibshirani, 1997), in which 5-fold cross-validation was performed to determine features with maximal correlation to survival outcomes (Figure S5). This led to the identification of two sets of three different features, one set for OS and the second for DFS.

Construction of radiomic signature

The prediction models were built using the Cox proportional hazards model using the corresponding features for OS and DFS, respectively. The radiomic signatures were calculated based on the coefficients of the corresponding features in the models. For visualization purposes, OS and DFS radiomic nomograms were constructed based on the two radiomic signatures. These can serve as low-cost, non-invasive means to predict the risk of death or relapse for HGSOC patients.

Prognostic performance evaluation

The predictive performance of the survival models was evaluated in the training cohort and further verified in both the internal and external validation cohorts using the concordance index (C-index) (Harrell et al., 1996). In addition, the receiver operation characteristics (ROC) curves of 2-6 years OS and 1–5 years DFS were plotted for the three datasets. Patients were divided into a high-risk group or low-risk group of corresponding specific years for both OS and DFS according to the cut-off risk score determined by these ROC curves. The optimal cut-off values of each stratification were determined using the Youden index (Hu et al., 2019) where the sum of sensitivity and specificity was maximized in the training dataset. The Kaplan-Meier survival analysis and Log rank test were used to compare the differences between the survival curves of these paired groups (van Timmeren et al., 2017).

Comparison of the feature selection methods

Comparative experiments between our FAMUS method and other commonly used methods for feature selection were conducted to reveal the importance of SASV features in radiomics-based prediction of clinical outcomes of HGSOC. To evaluate the performance of survival models using features pre-scanned by FAMUS, 200 experiment repetitions were performed. Each time, LASSO, sequential forward floating selection (SFFS) (Tan et al., 2014), ReliefF (Pupo et al., 2013), or mRMR (X Fau et al., 2021) were used to select features from FAMUS (i.e., SCC + FAMUS + LASSO, SCC + FAMUS + SFFS, SCC + FAMUS + ReliefF, and SCC + FAMUS + mRMR) or from SCC (i.e., SCC + LASSO, SCC + SFFS, SCC + ReliefF, and SCC + mRMR) directly using the same samples from the training dataset. Three-quarters of training datasets were sampled in each experiment. The C-Index was used to evaluate the performance of survival models

based on these methods, and the IQRs of the recorded C-Index of each method was displayed in box-plots to compare these methods intuitively. Average Hamming Distance (AHD) was used to assess the stability of the results of feature selection, and a lower AHD indicated a more stable feature selection result (Dunne et al., 2002). A paired *t*-test was performed to evaluate the improvement margin of FAMUS over other methods. Additionally, the top two features with the highest and lowest frequency in FAMUS were plotted to interpret how FAMUS worked.

QUANTIFICATION AND STATISTICAL ANALYSIS

The 95% confident interval (CI) of C-index was calculated by 1,000 bootstrap resamples. The differences in patient survival characteristics in the three cohorts were compared using a Log rank test. The differences in patient clinical characteristics in the three cohorts were compared by using the Kruskal-Wallis test or Chi-square test. Chi-square test, nomogram models and their calibration curves were performed using R (version 3.5.1, http://www.R-project.org). All other statistical analyses were performed with Python (version 3.6.4, https://www.python.org).