2 Materials and methods

2.1 Acquisition of gene expression and clinical data

The mRNA expression profiles and clinical information of OvCa tissue samples were obtained from The Cancer Genome Atlas (TCGA/TCGA-OV; https://portal.gdc.cancer.gov) and Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/; GSE26712). Given the privilege of sample size, the TCGA-OV cohort was further designated to build the risk signature. In addition, two external validation cohorts of OvCa cases with RNA sequencing and clinical data were downloaded from the International Cancer Genome Consortium (ICGC; https://dcc.icgc.org/projects/OV-AU) and another dataset of GEO (GSE49997). The samples with survival time less than 30 days and loss to follow-up have been deleted. The expression data of each gene from TCGA-OV, ICGC, and GEO dataset was normalized by the "SCALE" function to avoid deviations caused by different sequencing platforms ahead of validation. Finally, a total of 785 ER stress-related genes extracted from the GeneCards database (version 5.2) (https://www.genecards.org/; with a relevance score ≥7; Table 1) were prepared for subsequent analysis.

2.2 Construction of an ER stress-related risk signature

To construct an ER stress-related signature for survival prediction, univariate Cox analysis was firstly applied to screen out the prognosis-related genes in TCGA-OV and GSE26712 cohorts ("survminer" R package; version 0.4.9; with a cutoff of $P<0.05$). After intersection, the prognostic genes were aligned in the ER stress-related gene set. By employing the "glmnet" R packages (version 4.1-3), eight alignment genes were involved in the least absolute shrinkage and selection operator (LASSO) regression model to construct a risk signature. Subsequently, a seven-gene ER stress-related risk signature was generated after removing the gene named paxillin (*PXN*). The risk score for each sample was calculated as following formula: Risk score=$\sum_i^n Xi \times Yi$, where Xi is the relative expression value of each selected gene and Yi is the coefficient obtained from LASSO analysis.

Based on the median score, the same parameters were employed to classify patients in the training and validation cohorts into low- and high-risk groups. Principal component analysis (PCA) and *t*-distributed stochastic neighbor embedding (*t*-SNE) analysis in the "tinyarray" R package (version 2.2.7) were utilized to unveil the different clusters of the two groups. We then used the "survival" R package (version 3.3-1) to draw a Kaplan-Meier (K-M) plot to evaluate the overall survival (OS) between the two groups. The time-dependent receiver operating characteristic (ROC; "pROC" R package; version 1.18.0) curve and the area under the curve (AUC) over time were used to assess the sensitivity and specificity of the gene signature.

2.3 Identification of independent risk factors and construction of nomograms

To identify the independent prognostic value of this signature, we performed univariate and

multivariate regression analysis including the seven-gene risk signature and common clinical parameters. Data of regression analysis were presented by forest plots ("forestplot" R package; version 2.0.1). Furthermore, the ER stress-related signature was included in the construction of the predictive nomogram ("rms" R package; version 6.2-0). The calibration curves were used for internal verification of nomograms.

2.4 Validation of the ER stress-related risk signature

The median risk score was adopted to divide the patients of two external cohorts into low- and high- risk groups as described before. Then, the K-M and time-dependent ROC curves were used to test the predictive performance of the ER stress-related risk model. The independent prognostic value of this risk model was further determined by the regression analysis performed on two validation cohorts. Nomograms were constructed again to realize the predictive value of this signature.

2.5 Functional enrichment analysis

To reveal biological functions and the potential pathways of ER stress-related genes, Gene Ontology analysis (GO; http://www.geneontology.org) and Kyoto Encyclopedia of Genes and Genomes analysis (KEGG; http://www.kegg.jp) with the "clusterProfiler" package (version 4.2.2) in R software were also utilized in this study. The top five GO terms as well as KEGG-enriched pathways were visualized based on R package "ggplot2" (version 3.3.5). The gene set enrichment analysis (GSEA) approach was supplemented for less efficiency of KEGG. In addition, the relationship between risk model and canonical ER stress pathway gene was demonstrated by correlation analysis based on Kruskal's test.

2.6 Immune infiltrating profiles analysis

CIBERSORT algorithm (https://cibersort.stanford.edu/) was applied to clarify the correlation of ER stress-related risk signature and immune infiltration. Heatmap was preformed to visualize the distribution of immune cells in each sample of TCGA-OV cohort; and boxplots were drawn to illustrate the difference of infiltrating profiles between low- and high-risk groups. The correlation between *CD8* and *TAP1* expression was performed using "ggpubr" R package (version 0.4.0) to illustrate the special role of TAP1 in pMHC-I formation.

2.7 Statistical analysis

R software (version 4.1.2) was mainly used as statistical analysis tools in this study. Pearson's χ^2 test was employed to compare the categorical variables. The K-M curves with a two-sided log-rank test were applied to compare the OS of patients between groups. Univariate and multivariate Cox regression analysis, nomogram model, and time-dependent ROC curve analysis were used to evaluate the prognostic value of the ER stress-related signature. $P<0.05$ was considered as statistically significant.