

RESEARCH

Open Access



Machine learning-based prediction of survival prognosis in cervical cancer

Dongyan Ding^{1,2†}, Tingyuan Lang^{1,2,3*†}, Dongling Zou², Jiawei Tan⁴, Jia Chen⁴, Lei Zhou^{5,6,7}, Dong Wang², Rong Li², Yunzhe Li^{1,2}, Jingshu Liu^{1,2}, Cui Ma⁸ and Qi Zhou^{1,2,3*}

*Correspondence:
michaellang2009@163.com;
cqzl_zq@163.com

[†]Dongyan Ding and
Tingyuan Lang contributed
equally to this work.

² Department of Gynecologic
Oncology, School
of Medicine, Chongqing
University Cancer Hospital,
, Chongqing University,
Chongqing 400030, People's
Republic of China
Full list of author information
is available at the end of the
article

Abstract

Background: Accurately forecasting the prognosis could improve cervical cancer management, however, the currently used clinical features are difficult to provide enough information. The aim of this study is to improve forecasting capability by developing a miRNAs-based machine learning survival prediction model.

Results: The expression characteristics of miRNAs were chosen as features for model development. The cervical cancer miRNA expression data was obtained from The Cancer Genome Atlas database. Preprocessing, including unquantified data removal, missing value imputation, samples normalization, log transformation, and feature scaling, was performed. In total, 42 survival-related miRNAs were identified by Cox Proportional-Hazards analysis. The patients were optimally clustered into four groups with three different 5-years survival outcome ($\geq 90\%$, $\approx 65\%$, $\leq 40\%$) by K-means clustering algorithm base on top 10 survival-related miRNAs. According to the K-means clustering result, a prediction model with high performance was established. The pathways analysis indicated that the miRNAs used play roles involved in the regulation of cancer stem cells.

Conclusion: A miRNAs-based machine learning cervical cancer survival prediction model was developed that robustly stratifies cervical cancer patients into high survival rate (5-years survival rate $\geq 90\%$), moderate survival rate (5-years survival rate $\approx 65\%$), and low survival rate (5-years survival rate $\leq 40\%$).

Keywords: Cervical cancer, miRNAs, Machine learning, Survival prediction, Support-vector machines

Background

Cervical cancer is the main cause of women deaths worldwide which accounts for more than 520,000 new cases and 260,000 deaths each year [1, 2]. Although vaccines against the prime carcinogenic human papilloma virus (HPV) types are available commercially, the proportion of women receiving the vaccine is still low, especially in developing countries [3, 4]. Furthermore, despite effective treatment of early cervical cancer with surgery and radiation therapy, late cervical cancer is usually uncontrollable [5, 6].



© The Author(s) 2021. This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

Survival prediction after first diagnosis is important for both disease specialist and patients or their family members. First, as the survival ability of the cancer patients largely depends on the malignancy of the cancer cells, accurately forecasting the prognosis would be helpful for estimating the degree of malignancy and the time point of disease progression [7, 8]. On the other hand, patients and the families can set appropriate goals base on the accurate survival prediction. As the result, the timely prevention and treatment would be made and the worse treatment decision, such as over-treatment or late palliative care, would be effectively avoided.

However, accurate prediction of survival of cervical cancer patients is still challenging due to the heterogeneity of the cancer cells. In general, the cervical cancer patients were stratified into different groups base on cancer staging systems, TNM classification [9–11], for example. However, the molecular features have been rarely considered in such staging systems that numbers of subtypes of patients with different survival outcomes would be existed in one specific TNM stage. Furthermore, clinical features, such as TNM stage, could not provide enough information for survival prediction. For example, we investigated the correlation between several clinical features (including age at initial pathologic diagnosis, age began smoking, neoplasm pathologic margin involved text, neoplasm pathologic margin involved type, height, histological type, race, clinical stage, HPV infection) and the survival of cervical cancer patients included in The Cancer Genome Atlas (TCGA) dataset and found that only clinical stage is relevant, however, its discrimination ability is still not enough for accurate prediction (Log-rank $p=0.012$) (Fig. 1a and b). Thus, novel prediction strategies were urgently needed to be developed.

Molecular features (such as gene or noncoding RNA expression levels, gene mutation, copy number variation, etc.) imply substantial information about cancer cells, including malignant level, metastasis ability and therapeutic sensitivity, etc. [12]. Several cancers (colorectal, breast, and cervical cancers, for example) have been stratified into subtypes base on the molecular profiles provided by rapidly developing database of cancer molecular information, such as TCGA [13–15]. Thus, development of molecular features-based prediction model keeps the promise for improving the accuracy of cancer survival prediction model. Furthermore, as a part of artificial intelligence, machine learning (ML) provides a solution for accuracy improvement of cancer survival prediction model. Machine learning is a process for analysis of big data, that was characterized as learning from mistakes and experiences [16, 17]. Several machine learning models, such as support vector machines (SVM), have been widely used for development of prediction model base on electronic medical record, images as well as molecular features of cancer cells [18–22].

Thus, the objective of this study is to develop a novel molecular features-based machine learning cervical cancer survival prediction model (CCSPM) with high performance. MiRNAs were chosen as features and Cox-PH, K-means clustering and SVM algorithms were used for survival-related features identification, features-based objectives stratification, and prediction model development, respectively. The results of this study would improve the forecasting capacity of CCSPM and be helpful for cervical cancer management.

Characteristics	Number of patients		P value
	<=3 years	>3 years	
Age at initial pathologic diagnosis			
Age <= 35	7	12	
35 < Age <= 60	33	60	0.74
Age > 60	15	20	
Age began smoking			
Age <= 30	13	24	
Age > 30	2	2	0.97
Neoplasm pathologic margin involved text			
Margins free of tumor	3	14	
Others	3	7	0.79
Neoplasm pathologic margin involved type			
Macroscopic and microscopic parametrial involvement	2	1	
Other location, specify	6	21	0.25
Positive bladder and vaginal margin	2	6	
Height (cm)			
< 150	3	3	
150 - 160	15	29	
160 - 170	21	42	0.85
>= 170	2	5	
Histological type			
Cervical squamous cell carcinoma	47	76	
Endocervical adenocarcinoma of the usual type	0	2	
Endocervical type of adenocarcinoma	3	6	0.85
Mucinous adenocarcinoma of endocervical type	4	6	
Others	1	2	
Race			
American indian or alaska native	2	0	
Asian	2	2	
Black or african american	5	11	0.24
Native hawaiian or other pacific islander	1	0	
White	39	66	
Clinical stage			
Stage I - IIa	27	60	
Stage IIb - IVb	28	31	0.07
HPV infection			
Hpv16	26	58	
Hpv18	5	8	0.24
Others	19	22	

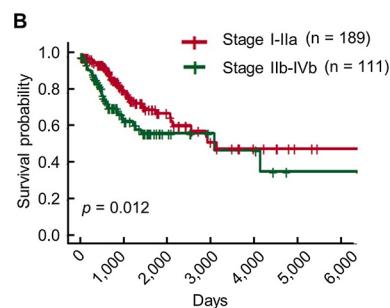


Fig. 1 The relationship between clinical features and survival of cervical cancer. **a** Chi-square analysis of the association between clinical features and 3-years survival of cervical cancer patients. **b** Kaplan–meier analysis of the association between clinical stage and the survival of cervical cancer. Chi-square was performed using chisq.test() function in "stats" (version 3.6.2) package. Kaplan–meier analysis was performed using Surv() and survfit() function in "survival" (version 3.2–10) package. The Kaplan–meier plot was performed using ggsurvplot() function in "survminer" (version 0.4.9). RStudio (version 3.6.1, RStudio, Inc.) is used

Results

Insufficient discriminative ability of clinical features for development of cervical cancer survival prediction model

Numbers of features with high discriminative ability is essential for development of a prediction model [23]. To develop the survival prediction model for cervical cancer, we investigated whether clinical features possess high discriminative ability for stratifying cervical cancer patients with different survival outcome. The information of clinical features and the survival of cervical cancer patients was downloaded from TCGA database [15]. The summary of the clinical information was given in Table 1. The cervical cancer patients were stratified base on 3-years survival outcome and the

Table 1 Summary of the clinical information of cervical cancer patients included in TCGA database

Characteristics	Cohort (n=307)	
	No.	%
<i>Age at initial pathologic diagnosis</i>		
Age <= 35	56	18.24
35 < Age <= 60	192	62.54
Age > 60	59	19.22
NA	0	0.00
<i>Age began smoking</i>		
Age <= 30	75	24.43
Age > 30	10	3.26
NA	222	72.31
<i>Neoplasm pathologic margin involved text</i>		
Margins free of tumor	23	7.49
Others	17	5.54
NA	267	86.97
<i>Neoplasm pathologic margin involved type</i>		
Macroscopic and microscopic parametrial involvement	10	3.26
Other location, specify	40	13.03
Positive bladder and vaginal margin	13	4.23
NA	244	79.48
<i>Survival (years)</i>		
<= 3	55	17.92
> 3	92	29.97
Loss to follow up < 3	160	52.12
<i>Height (cm)</i>		
< 150	15	4.89
150 - 160	119	38.76
160 - 170	111	36.16
>= 170	19	6.19
NA	43	14.01
<i>Histological type</i>		
Adenosquamous	6	1.95
Cervical squamous cell carcinoma	254	82.74
Endocervical adenocarcinoma of the usual type	6	1.95
Endocervical type of adenocarcinoma	21	6.84
Endometrioid adenocarcinoma of endocervix	3	0.98
Mucinous adenocarcinoma of endocervical type	17	5.54
NA	0	0.00
<i>Race</i>		
American indian or alaska native	8	2.61
Asian	20	6.51
Black or african american	30	9.77
Native hawaiian or other pacific islander	2	0.65
White	211	68.73
NA	36	11.73
<i>Clinical stage</i>		
Stage I - IIa	189	61.56
Stage IIb - IVb	111	36.16
NA	7	2.28
<i>HPV infection</i>		

Table 1 (continued)

Characteristics	Cohort (n=307)	
	No.	%
Hpv16	172	56.03
Hpv18	39	12.70
Others	73	23.78
NA	23	7.49

NA: Not available, HPV: human papillomavirus

clinical features were grouped as shown in Fig. 1a. Chi-square analysis was performed to determine whether clinical features were correlated with 3-years survival outcome. Unfortunately, no clinical features were found to be correlated with 3-years survival of cervical cancer base on Chi-square analysis (Fig. 1a), while, clinical stage has shown its potential ($p = 0.07$). We next performed Kaplan–meier analysis to confirm the link of clinical stage with cervical cancer survival. As shown in Fig. 1b, although the two groups of patients stratified by clinical stage exhibited statistically different survival outcome, its discriminative ability is still not enough for development of a survival prediction model (Log-rank p value = 0.012). These results suggested that clinical features, including clinical stage, could not provide enough information for development of CCSPM.

Data preprocessing of TCGA miRNAs expression data

We next investigate whether the expression of microRNAs (miRNAs) could be served as features for development of CCSPM. The miRNA expression data from TCGA database was used in this study [15]. The data preprocessing was first performed; data preprocessing plays important roles for statistical analysis of big data, including elimination of the impact from unquantified samples and features, missing values and outliers, reduction of batch-effects and experimental deviation, and normalization of the range of independent features, etc. [24].

The TCGA cervical cancer miRNA expression data (RPKM) was downloaded, in which 542 miRNAs in 312 cervical cancer samples were included. The reduced number of miRNAs is the result of the fact that only a certain number of miRNAs express in a specific cell [25]. The workflow of data preprocessing was given in Fig. 2a. The 2 metastasis and 3 normal control samples were first removed from analysis. For missing values imputation, two independent steps were performed as both batch effects and subtypes of cancer samples derived from heterogeneity of individuals should be considered. KNN missing value imputation was performed in this study; KNN is a non-parametric classification method in which K nearest neighbors of the subject were determined by physical units [26]. If the sample size is not big enough to contain K certain subtype samples in one batch, this study for example, when the batch-effects were existed in the data, K nearest neighbors are K samples in one specific batch and when the batch-effects were eliminated, K nearest neighbors are K samples in one specific molecular subtype. Thus, we performed two independent steps for missing values imputation. We first imputed the missing values with average of K nearest neighbors in one batch by KNN imputation

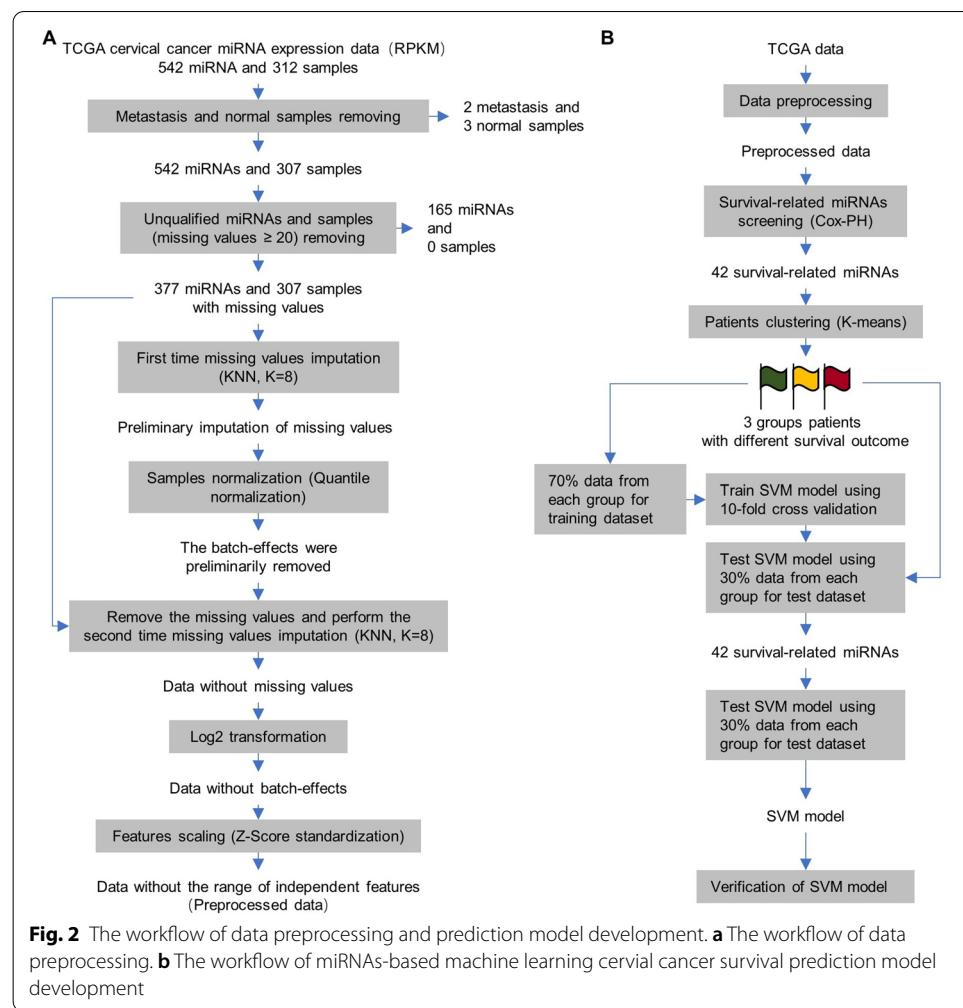


Fig. 2 The workflow of data preprocessing and prediction model development. **a** The workflow of data preprocessing. **b** The workflow of miRNAs-based machine learning cervical cancer survival prediction model development

method for subsequent sample normalization. After sample normalization by quantile normalization algorithm (the batch-effects were removed) [27], we removed the missing values and replaced them by KNN again with average of K nearest neighbors in one subtype. The data was finally processed by log2 transformation [28] and feature scaling (Z-score algorithm) [29, 30]. The miRNAs expression profiles derived from data after preprocessing was given in Additional file 1: Fig. 1.

Survival-related miRNAs identification

As shown in workflow of whole study (Fig. 2b), To develop the survival prediction model for cervical cancer, the features with high discriminative ability for cervical cancer survival are needed to be identified. MiRNA features were chosen as their important roles in cellular regulation and relative cost-effective for laboratory test. Cox-PH hazards model was used in this study. In total, 42 survival-related miRNAs were identified with log-rank p value less than 0.05; 23 and 19 miRNAs were positively and negatively correlated with survival ability of cervical cancer patients, respectively (Additional file 2: Table 1). The heatmap was given for exhibit the expression profiles of the survival-related miRNAs in tumor samples of patients (Additional file 1:

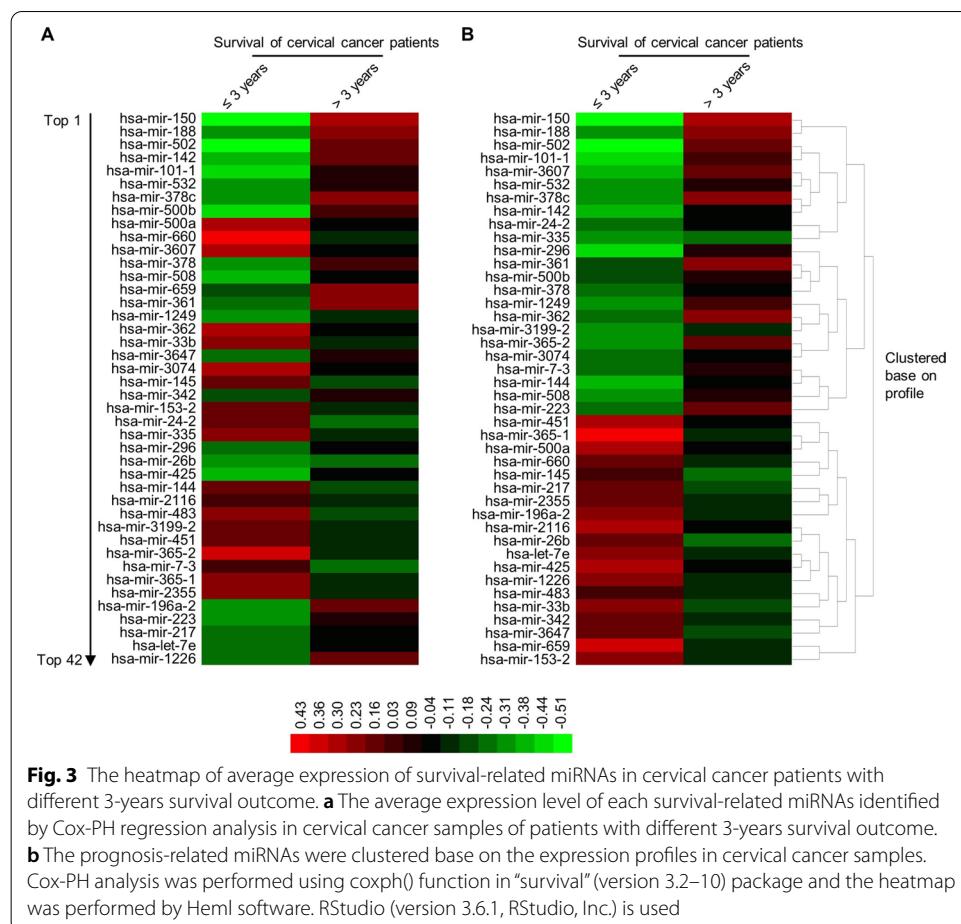


Fig. 3 The heatmap of average expression of survival-related miRNAs in cervical cancer patients with different 3-years survival outcome. **a** The average expression level of each survival-related miRNAs identified by Cox-PH regression analysis in cervical cancer samples of patients with different 3-years survival outcome. **b** The prognosis-related miRNAs were clustered base on the expression profiles in cervical cancer samples. Cox-PH analysis was performed using coxph() function in "survival" (version 3.2–10) package and the heatmap was performed by Heml software. RStudio (version 3.6.1, RStudio, Inc.) is used

Fig. 2), and the result show that the patients were clustered base on regular Euclidean distances derived from the expression profiles of the miRNAs. The Kaplan–meier survival curves were subsequently plotted for visualization of the discriminative ability of these miRNAs (Additional file 1: Fig. 3). Finally, the average expression of survival-related miRNAs in tumor samples of patients with different 3-years survival outcome was presented in (Fig. 3). However, although the survival-related miRNAs were identified, as the different discriminative ability of the features, the combination of these features should be optimized for prediction model development through K-means clustering analysis.

K-means clustering of cervical cancer patients with survival-related miRNAs

Next, the patients were stratified by K-means clustering algorithm with survival-related miRNAs; the grouped patients will be subsequently used for training prediction model. K-means clustering is a widely used machine learning program to partition n non-prelabeled observations into K clusters base on the characteristics of the features of the observations [31, 32]. In this study, we preformed the program when K=2 to 4, and the number of the features were optimized. As shown in Fig. 4, Kaplan–meier curves showed that the patients with different survival outcome were successfully separated by

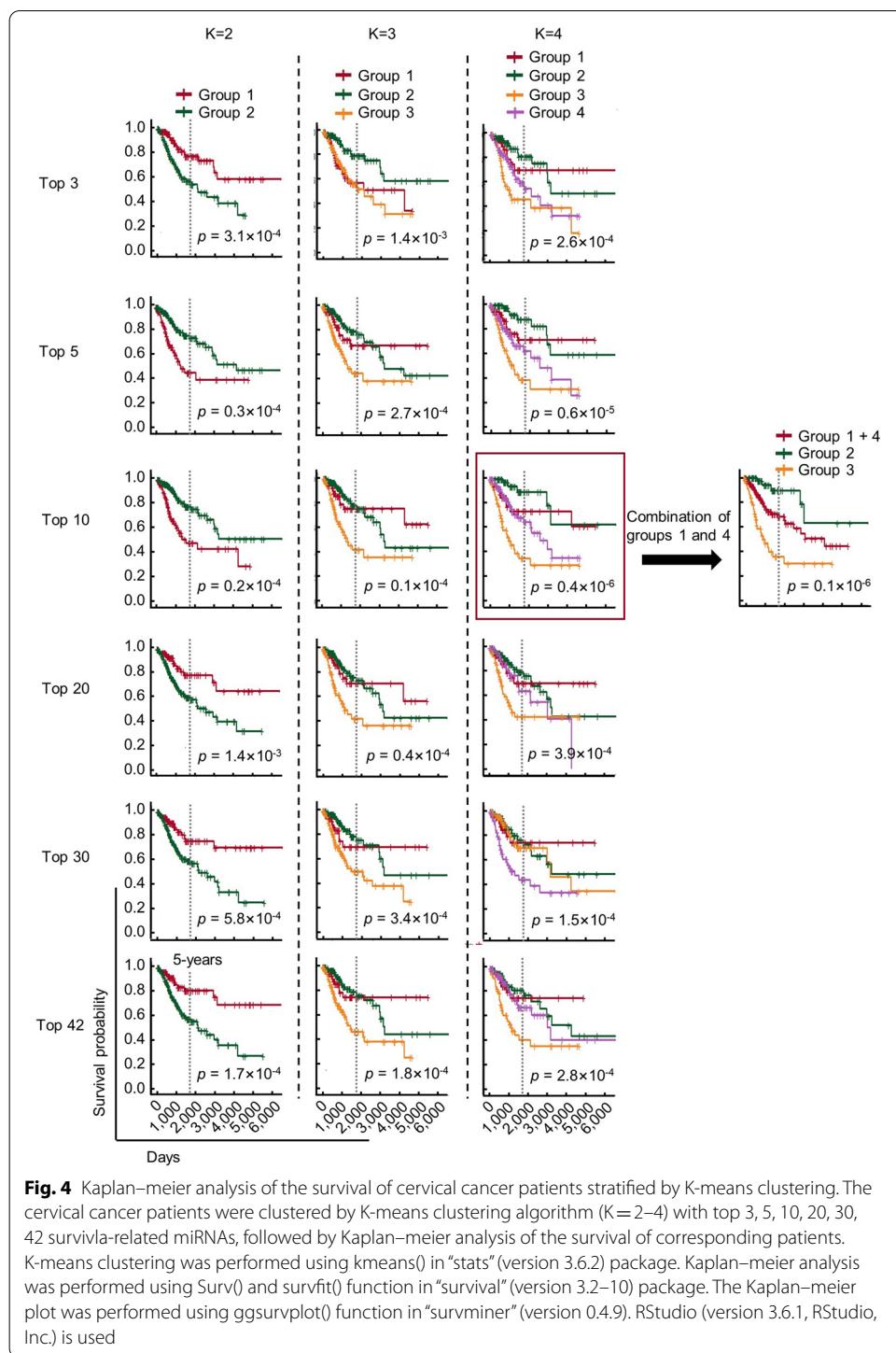
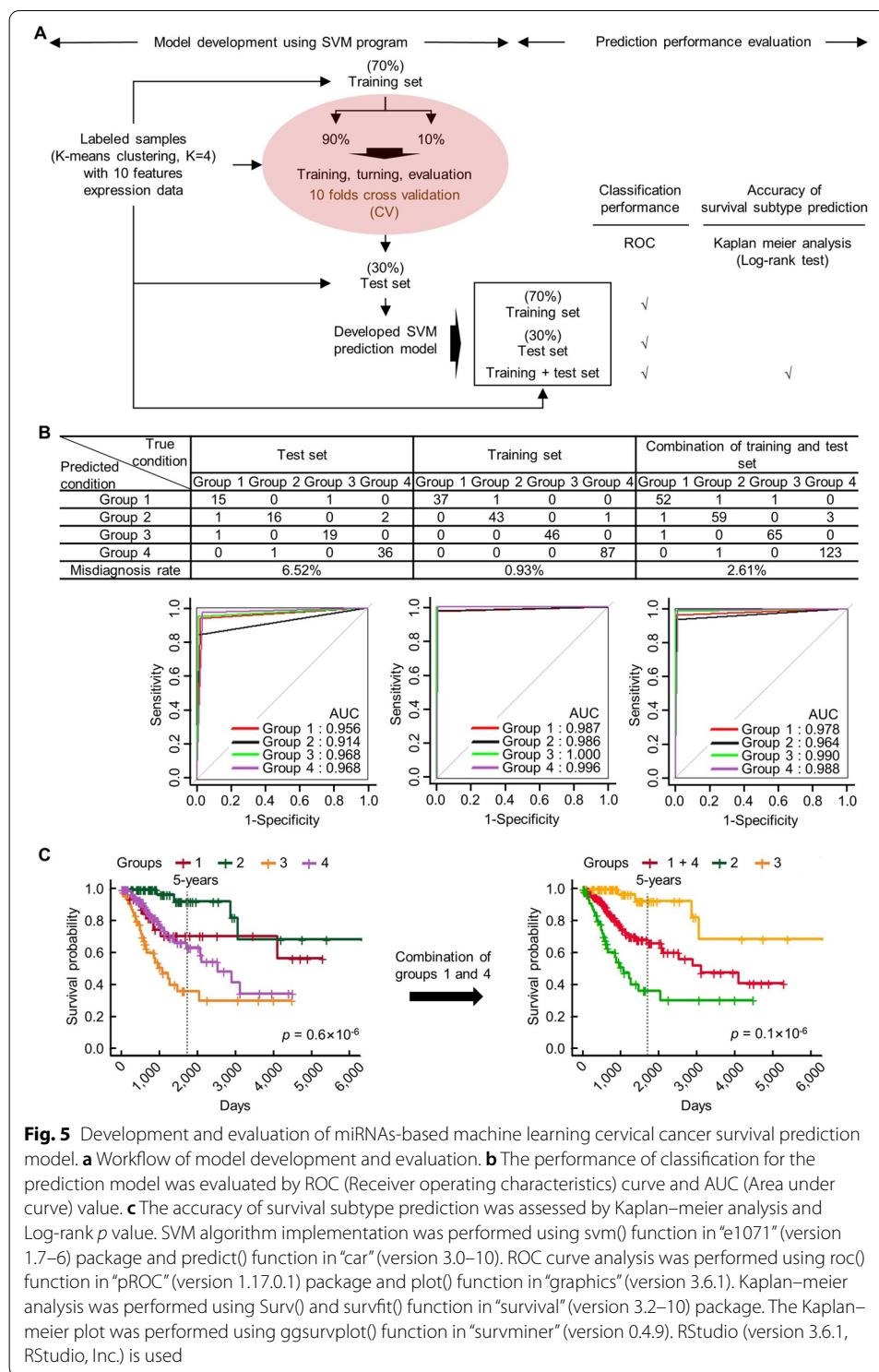
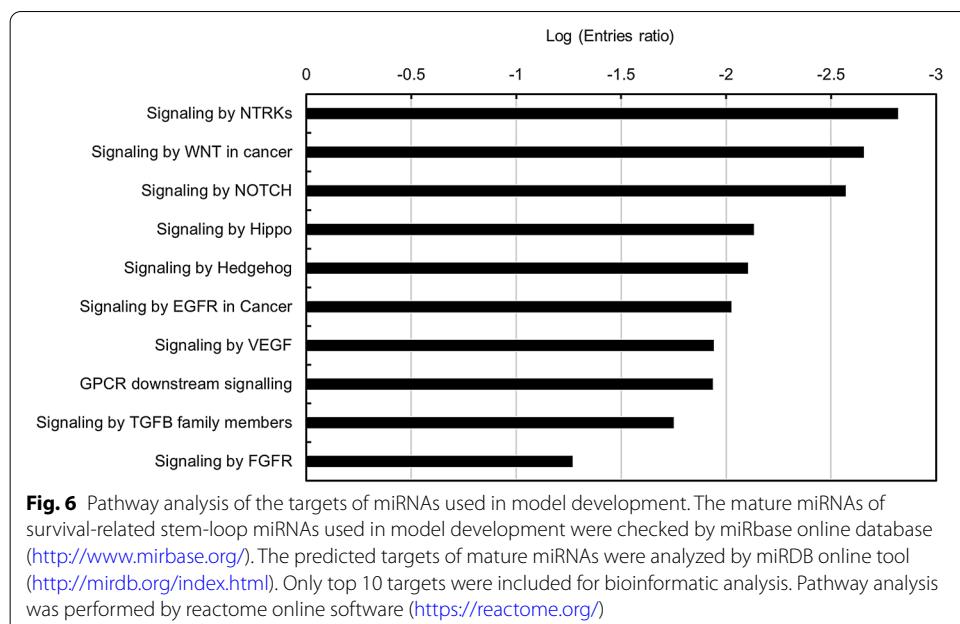


Fig. 4 Kaplan–meier analysis of the survival of cervical cancer patients stratified by K-means clustering. The cervical cancer patients were clustered by K-means clustering algorithm ($K=2\text{--}4$) with top 3, 5, 10, 20, 30, 42 survival-related miRNAs, followed by Kaplan–meier analysis of the survival of corresponding patients. K-means clustering was performed using `kmeans()` in "stats" (version 3.6.2) package. Kaplan–meier analysis was performed using `Surv()` and `survfit()` function in "survival" (version 3.2–10) package. The Kaplan–meier plot was performed using `ggsurvplot()` function in "survminer" (version 0.4.9). RStudio (version 3.6.1, RStudio, Inc.) is used

top 3, 5, 10, 20, 30 and all 42 miRNAs when $K=2$. Notably, when $K=4$, the program successfully separated the patients into four groups with three obviously different survival outcome base on top 10 survival-related miRNAs features. This result indicates that, theoretically, a prediction model with the ability to stratify patients into four groups with three different survival outcome (5-years survival rate $\geq 90\%$ (group 2), $\approx 65\%$ (group 1



and 4) and $\leq 40\%$ (group 3)) could be developed. Consideration of the biggest usefulness, this parameter (Top 10 miRNAs, $K=4$) was used for prediction model development.



SVM prediction model development

Next, the K-means clustering-derived of grouped patients and the related expression data of top 10 survival-related miRNAs were subjected into SVM program for prediction model training and development. SVM is one of the most powerful prediction methods for classification- or regression-aimed data analysis base on statistical learning frameworks and Vapnik–Chervonenkis (VC) theory [33]. In this study, a 7/3 split was used for generation of training and test data and tenfold cross-validation (CV) [34] was chosen. The workflow of SVM model development and evaluation was given in Fig. 5a.

After model development, the area under the operating characteristic curve (ROC) (AUC) was used to evaluate the discriminative ability of the model. As shown in Fig. 5b, the model exhibited high performance: AUC value = 0.956 (group 1), 0.914 (group 2), 0.968 (group 3), 0.968 (group 4) for test set; 0.987 (group 1), 0.986 (group 2), 1.000 (group 3), 0.996 (group 4) for training set; 0.978 (group 1), 0.964 (group 2), 0.990 (group 3), 0.988 (group 4) for whole set, and the misdiagnosis rate was 6.52% (test set), 0.93% (training set), 2.61% (whole set), respectively. Furthermore, the groups of patients predicted by the SVM model exhibited the similar survival outcome as the patients clustered by K-means algorithm (Fig. 5c), which confirmed the accuracy of the SVM prediction model. Collectively, a miRNAs-based ML CCSPM that stratifies cervical cancer patients into high survival rate (5-years survival rate $\geq 90\%$), moderate survival rate (5-years survival rate $\approx 65\%$) and low survival rate (5-years survival rate $\leq 40\%$) was developed.

Pathway analysis of the targets of miRNAs used in the model development

To understand the mechanisms underlying the miRNAs that served as the features for ML CCSPM, a pathway analysis was performed with the targets of these miRNAs. As shown in Fig. 6, the targets impacted the pathways involved in cancer stem cells (CSCs). It had been recognized that CSCs are the root of cancer initiation, progression, drug

Table 2 Comparation of SVM model developed in this study and Zhao model

Models	Groups stratified	5-years survival rate of each group	Log-rank p value	References
This study	3	≥90%, ≈ 65% and ≤ 40%	0.1×10^{-6}	
Zhao et al.	2	≈80% and 50%	<0.001	37

resistance, that leads to treatment failure [35]. This result indicated that the mechanism underlying the miRNAs selected as the features for ML CCSPM model is their roles in CSCs regulation.

Discussion

In this study, we developed a CCSPM base on TCGA database and machine learning technology which successfully stratified the cervical cancer patients into three groups (high survival ability, moderate survival ability, and low survival ability with 5-years survival rate $\geq 90\%$, $\approx 65\%$ and $\leq 40\%$).

Preprocessing is important for statistical analysis of biological data; several challenges are existed in biological samples-derived big data, including missing values, batch-effects and different ranges of independent features, etc. [24]. Sample normalization and feature scaling are normally necessary for data preprocessing, which were performed for elimination of batch-effects and normalizing the range of independent feature, respectively; quantile and Z-score algorithms were well-known and wildly-used methods for these two steps [27, 29, 30]. KNN methods are most popular strategy for missing values imputation [26], while, for cancer research, both batch-effects and subtypes of cancer samples are needed to be considered; the later one is usually ignored in studies. Cancer heterogeneity refers to the existence of subpopulations of tumor cells with different genotypes and phenotypes; this heterogeneity may exist in the same tumor or tumors from different patients [36]. Thus, ideally, the missing values should be replaced by the average values derived from same subtypes of tumor samples in certain batch. However, as the high cost of omics-data extraction, the number of patients included in the database is usually not enough. Therefore, two steps missing values imputation were adopted in this study; the missing values first were imputed by KNN method with average of values in certain batch for subsequent normalization, and after normalization, the missing values were imputed by KNN method again with average of values in certain subtype of samples (Fig. 2a). The strategy employed in this study maximumly attenuate the disturbance of missing values.

Accurately forecasting the survival cancer patients are important for therapeutic decision. Currently, most molecular-based survival prediction model stratified the patients into two groups with different survival outcome [37, 38]. For example, Zhao and colleagues developed a five-genes prognostic model that stratifies the cervical cancer patients into two groups with 5-years survival rate $\approx 80\%$ and 50% , respectively [37], while, this result is not accurate enough for therapeutic decision making. In this study, the 10 miRNAs-based prediction model developed by SVM program could robustly stratify the cervical patients into three groups (5-years survival rate $\geq 90\%$, $\approx 65\%$ and $\leq 40\%$), which significantly improves the usefulness of the model (Table 2).

In this study, we used K-means algorithm to stratify patients automatically base on miRNA expression characteristics, however, medical experience of researchers is helpful for the interpretation of the meaning of the clustering result. As shown in Fig. 4, it is difficult to decide that $K=4$ is the optimal clustering strategy according to the log-rank p value, while, although group 1 and group 4 possess distinct miRNA expression profiles, the patients show the similar 5-year survival outcome. Meanwhile, by this clustering strategy, the patients with high survival rate (group 2) and low survival rate (group 3) were stratified more accurately.

CSCs are the prime cause of cancer treatment as their congenital self-renewal capacity and enhanced metastasis, tumor-initiation and drug resistance abilities [35]. There are several pathways that play essential roles for CSCs maintenance; these pathways, such as Wnt, Hedgehog, Hippo and NOTCH etc. are also essential for normal stem cell regulation and development [39, 40]. The alteration of signaling involved in cancer microenvironment, including VEGF, FGFR, EGFR, GPCR, NTRKs and TGF β , etc. have also been found in CSCs [41–43]. Pathway analysis showed that the targets of the miRNAs used for CCSPM development significantly impacted these pathways (Fig. 6), indicating that the miRNAs associated with CSCs largely correlate with survival of cervical cancer patients and could be served as features for CCSPM development.

Conclusion

In summary, a miRNAs-based ML CCSPM was developed that robustly stratifies cervical cancer patients into high survival rate (5-years survival rate $\geq 90\%$), moderate survival rate (5-years survival rate $\approx 65\%$) and low survival rate (5-years survival rate $\leq 40\%$).

Methods

Datasets and data analysis

The miRNA expression reads per million of mapped reads per kilobase of transcript length (RPKM) data, which includes 542 miRNAs and 312 samples, and the related clinical information were downloaded by firehose online tools (file package name: gdac.broadinstitute.org_CESC.miRseq_Preprocess.Level_3.2016012800.0.0, miRNA file name: CESC.miRseq_RPKM.txt, clinical information file name: gdac.broadinstitute.org_CESC.Merge_Clinical.Level_1.2016012800.0.0) [15]. RPKM is a method for normalization of raw counts. As a specific cell expresses a certain number of miRNAs, not all 1046 miRNAs were included in the data. All data analysis with the exception of heatmap was performed in R [44] using RStudio (version 3.6.1, RStudio, Inc.).

Determination of the correlation between clinical features and survival of cervical cancer

The samples of metastasis tumor and normal tissues were first removed. The selected clinical features were grouped as shown in Fig. 2a and the patients in each group were divided based on 3-years survival (Fig. 2a). Chi-square test was used to determine the correlation between clinical features and 3-years survival of the patients. P values less than 0.05 were regarded as statistically significant. The potential features (P less than 0.1 by Chi-square analysis) were further confirmed by Kaplan–meier analysis. Chi-square

was performed using chisq.test() function in “stats” (version 3.6.2) package [44]. Kaplan–meier analysis was performed using Surv() and survfit() function in “survival” (version 3.2–10) package [45]. The Kaplan–meier plot was performed using ggsurvplot() function in “survminer” (version 0.4.9) [46].

Data preprocessing

The workflow of data preprocessing was presented in Fig. 3a. Unqualified samples and features removing, missing values imputation, samples normalization and features scaling were performed for miRNA expression data preprocessing. First, 2 metastasis and 3 normal samples were removed. Then, the miRNAs and samples with missing values ≥ 20 were removed. To perform sample normalization to eliminate the batch-effects, the missing values were preliminarily imputed by K-nearest neighbors (KNN) method, which impute the missing values by the average of K nearest neighbors, which were determined by Euclidean distances, in this study. KNN analysis was performed using knnImputation() function in “DMwR2” (version 0.0.2) [48].

Sample normalization was performed by quantile normalization method as described by Zhao’s publication [27]. The batch-effects were removed by this step. As not only batch-effects but also subtypes of heterogeneous tumor cells were considered for missing values imputation, the missing values were removed and replaced again by KNN method with normalized data. After log2 transformation, the feature scaling was performed by Z-score method. The preprocessed data was provided in Additional file 3: Table 2. The source code of data preprocessing was provided in GitHub (<https://github.com/dingdongyan/New-CESC-2021>). Quantile normalization was performed using normalize.quantiles() function in “preprocessCore” (version 1.48.0) package [48]. Calculation of average and standard deviation for Z-score analysis were performed using mean() function in “base” (version 3.6.2) package [44] and sd() function in “stats” (version 3.6.2) package [44].

Survival-related miRNAs identification

Cox proportional hazard (Cox-PH) model was used to identification of survival-related miRNAs. Cox-PH analysis was performed using coxph() function in “survival” (version 3.2–10) package [45]. All parameters were default. *P* values less than 0.05 were regarded as statistically significant. The expression profiles of survival-related miRNAs in patients were presented as heatmap. The heatmap was performed by Heml software [49]. The source code was provided in GitHub (<https://github.com/dingdongyan/New-CESC-2021>).

K-means clustering

R command, kmeans() was used to stratify the patients base on survival-related miRNAs. The expression data of top 3, 5, 10, 20, 30 and all miRNAs were input into K-means program. The parameters: centers = 2 to 4, inter.max = 10, nstart = 1, algorithm = Hartigan-Wong, trace = TRUE. The source code was provided in GitHub (<https://github.com/dingdongyan/New-CESC-2021>). K-means clustering was performed using kmeans() in “stats” (version 3.6.2) package [44].

Kaplan–meier analysis

Kaplan–meier analysis was used to calculate the survival rate of stratified patients and plot the survival curve. The analysis was performed by `Surv()` and `survfit()` function in “survival” (version 3.2–10) package [45] and the plot was performed by `ggsurvplot()` function in “survminer” (version 0.4.9) [4]. Log-rank p value for each analysis was given. The source code was provided in GitHub (<https://github.com/dingdongyan/New-CESC-2021>).

Supervised classification model development

The supervised classification model was developed by SVM algorithm with labeled samples and the expression data of survival-related miRNAs. The SVM model was developed by splitting the samples 70%/30% to training and held-out testing data. SVM algorithm implementation was performed using `svm()` function in “e1071” (version 1.7–6) package [50] and `predict()` function in “car” (version 3.0–10) [51]. The prime parameters: `type=C-classification`, `kernel=radial`, `gamma=0.05`, `cross=10`, `cost=5`, `scale=FALSE`. The source code was provided in GitHub (<https://github.com/dingdongyan/New-CESC-2021>).

ROC curve analysis

ROC curve analysis was performed using `roc()` function in “pROC” (version 1.17.0.1) package [52] and `plot()` function in “graphics” (version 3.6.1) [44]. The source code was provided in GitHub (<https://github.com/dingdongyan/New-CESC-2021>).

Bioinformatic analysis

The mature miRNAs of survival-related stem-loop miRNAs were checked by miRbase online database [53]. The predicted targets of mature miRNAs derived from survival-related miRNAs used for SVM model development were analyzed by miRDB online tool [54]. The top 10 targets were included for bioinformatic analysis. Pathway analysis was performed by reactome online software [55].

Abbreviations

HPV: Human papilloma virus; miRNA: MicroRNA; lncRNA: Long non-coding RNA; TCGA: The cancer genome atlas; ML: Machine learning; CCSPM: Cervical cancer survival prediction model; KNN: K-Nearest Neighbor; Cox-PH: Cox Proportional-Hazards; SVM: Support vector machine; ROC: Receiver operating characteristics; AUC: Area under the curve; CSCs: Cancer stem cells.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s12859-021-04261-x>.

[Additional file 1](#)

[Additional file 2](#)

[Additional file 3](#)

Acknowledgements

This study was funded by Chongqing Science & Technology Commission csc2019jscx-msxmX0174 and csc2019jscx-msxmX0106. We thank Professor Guangwu Tang from Chongqing research and Design Institute of Transportation for the idea of artificial intelligence. We thank Mr. Tingxiu Lang and Jin Chen from Shenzhen Forms Syntron Information Co., Ltd for the help in computer technology. We thank Professor Qing Zeng from Department of Public Health of Chongqing

Medical University for statistics idea. The results shown here are in part based upon data generated by the TCGA Research Network: <https://www.cancer.gov/tcga>.

Authors' contributions

DYD, TYL performed all the experiments. DLZ, JT interpreted the data. JC, LZ participated in statistical method development, DW, RL, YZL, JSL, CM performed data analysis involved in clinical information. TYL and DYD wrote the manuscript. TYL and QZ designed and supervised the study. All authors read and approved the final manuscript.

Funding

The study is funded by Chongqing Science & Technology Commission csc2019jscx-msxmX0174 and csc2019jscx-msxmX0106. The funder play no role in study design and execution.

Availability of data and materials

The datasets supporting the conclusions of this article are included within the article and its additional files. The codes are deposited in GitHub (<https://github.com/dingdongyan/New-CESC-2021>).

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The author reports no conflicts of interest in this work.

Author details

¹Key Laboratory of Biorheological Science and Technology (Chongqing University), Ministry of Education, Bioengineering College, Chongqing University, Chongqing 400044, People's Republic of China. ²Department of Gynecologic Oncology, School of Medicine, Chongqing University Cancer Hospital, Chongqing University, Chongqing 400030, People's Republic of China. ³Chongqing Key Laboratory of Translational Research for Cancer Metastasis and Individualized Treatment, School of Medicine, Chongqing University Cancer Hospital, Chongqing University, Chongqing 400030, People's Republic of China. ⁴School of Mathematics and Statistics, Changchun University of Technology, Changchun 130012, People's Republic of China. ⁵Singapore Eye Research Institute, The academia, 20 College Road, Discovery Tower Level 6, Singapore 169856, Singapore. ⁶Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore. ⁷Duke-NUS Medical School, Ophthalmology and Visual Sciences Academic Clinical Research Program, National University of Singapore, Singapore, Singapore. ⁸Department of Pediatric Hematology, First Hospital of Jilin University, Changchun 130023, Jilin, People's Republic of China.

Received: 23 December 2020 Accepted: 11 June 2021

Published online: 16 June 2021

References

- Paul A Cohen, Anjua Jhingran, Ana Oaknin, Lynette Denny. Cervical cancer. *The Lancet*, 2019;393.
- Canfell K, Kim JJ, Brisson M, Keane A, Simms KT, Caruana M, et al. Mortality impact of achieving WHO cervical cancer elimination targets: a comparative modelling analysis in 78 low-income and lower-middle-income countries. *Lancet*. 2020;395:591–603.
- Lei J, Ploner A, Elfström KM, Wang J, Roth A, Fang F, et al. HPV vaccination and the risk of invasive cervical cancer. *N Engl J Med*. 2020;383:1340–8.
- Williams EA, Newberg J, Williams KJ, Montesion M, Alexander BM, Lin DI, et al. Prevalence of High-Risk nonvaccine human papillomavirus types in advanced squamous cell carcinoma among individuals of African vs Non-African ancestry. *JAMA Netw Open*. 2021;4:e216481.
- Huang H, Feng YL, Wan T, Zhang YN, Cao XP, Huang YW, et al. Effectiveness of sequential chemoradiation vs concurrent chemoradiation or radiation alone in adjuvant treatment after hysterectomy for cervical cancer: the STARS phase 3 randomized clinical trial. *JAMA Oncol*. 2021;7:361–9.
- Arbyn M, Weiderpass E, Bruni L, de Sanjosé S, Saraiya M, Ferlay J, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *Lancet Glob Health*. 2020;8:e191–203.
- Yankeelov TE, Quaranta V, Evans KJ, Rericha EC. Toward a science of tumor forecasting for clinical oncology. *Cancer Res*. 2015;75:918–23.
- Ben-Hamo R, Jacob Berger A, Gavert N, Miller M, Pines G, Oren R, et al. Predicting and affecting response to cancer therapy based on pathway-level biomarkers. *Nat Commun*. 2020;11:3296.
- Knoth J, Pötter R, Jürgenliemk-Schulz IM, Haie-Meder C, Fokdal L, Sturdza A, et al. Clinical and imaging findings in cervical cancer and their impact on FIGO and TNM staging - an analysis from the EMBRACE study. *Gynecol Oncol*. 2020;159:136–41.
- Ryzhov A, Corbex M, Piñeros M, Barchuk A, Andreasyan D, Djanklich S, et al. Comparison of breast cancer and cervical cancer stage distributions in ten newly independent states of the former Soviet Union: a population-based study. *Lancet Oncol*. 2021;22:361–9.

11. Shin W, Ham TY, Park YR, Lim MC, Won YJ. Comparing survival outcomes for cervical cancer based on the 2014 and 2018 International Federation of Gynecology and Obstetrics staging systems. *Sci Rep.* 2021;11:6988.
12. Corces MR, Granja JM, Shams S, Louie BH, Seoane JA, Zhou W, et al. The chromatin accessibility landscape of primary human cancers. *Science.* 2018;362:eaav1898.
13. Network CGA. Comprehensive molecular characterization of human colon and rectal cancer. *Nature.* 2012;487:330–7.
14. Network CGA. Comprehensive molecular portraits of human breast tumours. *Nature.* 2012;490:61–70.
15. Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, Barretos Cancer Hospital, Baylor College of Medicine, Beckman Research Institute of City of Hope, et al. Integrated genomic and molecular characterization of cervical cancer. *Nature.* 2017;543:378–84.
16. Kc G, Bocci G, Verma S, Hassan M, Holmes J, Yang J, et al. A machine learning platform to estimate anti-SARS-CoV-2 activities. *Nat Mach Intell.* 2021. <https://doi.org/10.1038/s42256-021-00335-w>.
17. Schlick T, Portillo-Ledesma S. Biomolecular modeling thrives in the age of technology. *Nat Comput Sci.* 2021. <https://doi.org/10.1038/s43588-021-00060-9>.
18. Wulczyn E, Steiner DF, Moran M, Plass M, Reihs R, Tan F, et al. Interpretable survival prediction for colorectal cancer using deep learning. *NPJ Digit Med.* 2021;4:71.
19. Gangayag MD, Taib NA, Har YC, Lio P, Dhillon SK. Predicting factors for survival of breast cancer patients using machine learning techniques. *BMC Med Inform Decis Mak.* 2019;19:48.
20. Li J, Zhou Z, Dong J, Fu Y, Li Y, Luan Z, et al. Predicting breast cancer 5-year survival using machine learning: a systematic review. *PLoS One.* 2021;16:e0250370.
21. Nagy M, Radakovich N, Nazha A. Machine learning in oncology: what should clinicians know? *JCO Clin Cancer Inform.* 2020;4:799–810.
22. Huang S, Yang J, Fong S, Zhao Q. Artificial intelligence in cancer diagnosis and prognosis: opportunities and challenges. *Cancer Lett.* 2020;471:61–71.
23. Nichols JA, Herbert Chan HW, Baker MAB. Machine learning: applications of artificial intelligence to imaging and diagnosis. *Biophys Rev.* 2019;11:1111–8.
24. Heidari M, Mirniahari-kandehi S, Khuzani AZ, Danala G, Qiu Y, Zheng B. Improving the performance of CNN to predict the likelihood of COVID-19 using chest X-ray images with preprocessing algorithms. *Int J Med Inform.* 2020;144:104284.
25. Gebert LFR, MacRae IJ. Regulation of microRNA function in animals. *Nat Rev Mol Cell Biol.* 2019;20:21–37.
26. Cherif W. Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis. *Proced Comput Sci.* 2018;127:293–9.
27. Zhao Y, Wong L, Goh WWB. How to do quantile normalization correctly for gene expression data analyses. *Sci Rep.* 2020;10:15534.
28. Wang Q, Armenia J, Zhang C, Penson AV, Reznik E, Zhang L, et al. Unifying cancer and normal RNA sequencing data from different sources. *Sci Data.* 2018;5:180061.
29. Cancer Genome Atlas Research Network, Albert Einstein College of Medicine, Analytical Biological Services, Barretos Cancer Hospital, Baylor College of Medicine, Beckman Research Institute of City of Hope, et al. Integrated genomic and molecular characterization of cervical cancer. *Nature.* 2017;543:378–84.
30. Burdack J, Horst F, Giesselbach S, Hassan I, Daffner S, Schöllhorn WI. Systematic comparison of the influence of different data preprocessing methods on the performance of gait classifications using machine learning. *Front Bioeng Biotechnol.* 2020;8:260.
31. Altman N, Krzywinski M. Clustering. *Nat Methods.* 2017;14:545–6.
32. Cao XH, Stojkovic I, Obradovic Z. A robust data scaling algorithm to improve classification accuracies in biomedical data. *BMC Bioinformatics.* 2016;17:359.
33. Stafford IS, Kellermann M, Mossotter E, Beattie RM, MacArthur BD, Ennis S. A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. *NPJ Digit Med.* 2020;3:30.
34. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. *BMC Med Inform Decis Mak.* 2019;19:281.
35. Batlle E, Clevers H. Cancer stem cells revisited. *Nat Med.* 2017;23:1124–34.
36. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. *Nat Rev Clin Oncol.* 2018;15:81–94.
37. Zhao M, Huang W, Zou S, Shen Q, Zhu X. A five-genes-based prognostic signature for cervical cancer overall survival prediction. *Int J Genomics.* 2020;2020:8347639.
38. Chaudhary K, Poirion OB, Lu L, Garmire LX. Deep learning-based multi-omics integration robustly predicts survival in liver cancer. *Clin Cancer Res.* 2018;24:1248–59.
39. Yang L, Shi P, Zhao G, Xu J, Peng W, Zhang J, et al. Targeting cancer stem cell pathways for cancer therapy. *Signal Transduct Target Ther.* 2020;5:8.
40. Takebe N, Miele L, Harris PJ, Jeong W, Bando H, Kahn M, et al. Targeting Notch, Hedgehog, and Wnt pathways in cancer stem cells: clinical update. *Nat Rev Clin Oncol.* 2015;12:445–64.
41. Du J, Xu Y, Sasada S, Oo AOKK, Hassan G, Mahmud H, et al. Signaling inhibitors accelerate the conversion of mouse iPS cells into cancer stem cells in the tumor microenvironment. *Sci Rep.* 2020;10:9955.
42. Dorsam R, Gutkind J. G-protein-coupled receptors and cancer. *Nat Rev Cancer.* 2007;7:79–94.
43. Somwar R, Hofmann NE, Smith B, Odintsov I, Vojnic M, Linkov I, et al. NTRK kinase domain mutations in cancer variably impact sensitivity to type I and type II inhibitors. *Commun Biol.* 2020;3:776.
44. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2019. <https://www.R-project.org/>.
45. Terry M Therneau. A Package for Survival Analysis in R. New York; R package version 3.2–10. 2021. <https://CRAN.R-project.org/package=survival>.
46. Alboukadel Kassambara, Marcin Kosinski, Przemyslaw Biecek. survminer: Drawing Survival Curves using 'ggplot2'. R package version 0.4.9. 2021. <https://CRAN.R-project.org/package=survminer>.

47. Torgo, L. Data Mining with R, learning with case studies, 2nd edition Chapman and Hall/CRC.2016. R package version 0.0.2. <http://ltorgo.github.io/DMwR2>.
48. Ben Bolstad. preprocessCore: A collection of pre-processing functions. R package version 1.48.0. 2019. <https://github.com/bmbolstad/preprocessCore>.
49. Wankun Deng, Yongbo Wang, Zexian Liu, Han Cheng, Yu Xue. Heml: A Toolkit for Illustrating Heatmaps. PLoS One. 2014;9:e111988.
50. David Meyer, Evgenia Dimitriadou, Kurt Hornik, Andreas Weingessel, Friedrich Leisch. e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.7–6. 2021. <https://CRAN.R-project.org/package=e1071>.
51. John Fox, Sanford Weisberg. An {R} Companion to Applied Regression. Thousand Oaks CA: Sage. R package version 3.0–10. 2019. <https://socialsciences.mcmaster.ca/jfox/Books/Companion/>.
52. Xavier Robin, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics. 2011; 12:77. <http://www.biomedcentral.com/1471-2105/12/77>.
53. Kozomara A, Birgaoanu M, Griffiths-Jones S. miRBase: from microRNA sequences to function. Nucleic Acids Res. 2019;47:D155–62.
54. Chen Y, Wang X. miRDB: an online database for prediction of functional microRNA targets. Nucleic Acids Res. 2020;48:D127–31.
55. Jassal B, Matthews L, Viteri G, Gong C, Lorente P, Fabregat A, et al. The reactome pathway knowledgebase. Nucleic Acids Res. 2020;48:D498–503.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions

