Research letter

Oncology

Deep learning survival model for colorectal cancer patients (DeepCRC) with Asian clinical data compared with different theories

Wei Li[1], Shuye Lin[1], Yuqi He[2], Jinghui Wang[1,3], Yuanming Pan[1]

[1]Cancer Research Center, Beijing Chest Hospital, Capital Medical University, Beijing Tuberculosis and Thoracic Tumor Research Institute, Tongzhou District, Beijing, China
[2]Department of Gastroenterology, Beijing Chest Hospital, Capital Medical University, Tongzhou District, Beijing, China
[3]Department of Oncology, Beijing Chest Hospital, Capital Medical University, Tongzhou District, Beijing, China

Submitted: 18 September 2022; **Accepted:** 12 November 2022
Online publication: 13 January 2023

Arch Med Sci 2023; 19 (1): 264–269
DOI: https://doi.org/10.5114/aoms/156477

Corresponding authors:
Yuanming Pan PhD
Cancer Research Cente
Beijing Chest Hospital
Capital Medical University
Beijing Tuberculosis and Thoracic Tumor
Research Institute
Beijing 101149, China
Phone: +86-10-89509372
Fax: +86-10-89509372
E-mail: peter.f.pan@ hsc.pku.edu.cn

Prof. Jinghui Wang MD
Cancer Research Center
Beijing Chest Hospital
Capital Medical University
Beijing Tuberculosis and Thoracic Tumor
Research Institute
Beijing 101149, China
Phone: +86-10-89509372
Fax: +86-10-89509372
E-mail: jinghuiwang2006@163.com

Abstract

Introduction: Colorectal cancer (CRC) is the third most common cancer. Precise prediction of CRC patients' overall survival (OS) probability could offer advice on its treatment. Neural network (NN) is the first-class algorithm, but a consensus on which NN survival models are better has not been established yet. A predictive model on CRC using Asian data is also lacking.
Methods: We conducted 8 NN survival models of CRC (n = 416) with different theories and compared them using Asian data.
Results: DeepSurv performed best with a C-index value of 0.8300 in the training cohort and 0.7681 in the test cohort.
Conclusions: The deep learning survival model for CRC patients (DeepCRC) could predict CRC's OS accurately.

Key words: colorectal cancer, neural network, deep learning, predictive model.

Colorectal cancer (CRC) is the third most common cancer, accounting for about 10% annually diagnosed tumors worldwide, and it is the second leading cause of death from among all tumors [1, 2]. Given the impairment of quality of life from not only CRC itself but also the treatment's adverse effects, such as a stoma, it is pivotal to predict a patient's overall survival (OS).

American Joint Committee on Cancer (AJCC) TNM stage is a typical and extensively used reference for cancer prognosis. However, many studies have revealed that the survival of the same stage CRC patients varied, and a more precise staging system is needed [3–7]. Another choice is to use the Cox proportional hazard model (CPH). But the CPH is a semiparametric model, assuming that a patient's log-risk of an event (e.g., "death") is a linear combination of the patient's covariates, which might be too simplistic to handle time-to-event prediction in the real world [8, 9]. In this regard, some researchers began to set their sights on machine learning algorithms and even deep learning neural networks (NNs). NNs can improve prediction accuracy by discovering relevant features

of high complexity [8, 9]. There are 8 popular NN survival theories, such as DeepSurv and CoxCC (Cox case-control corresponding methods). However, no study has compared them yet. At the same time, though there have been some predictive models for CRC, they were mainly based on the CPH, traditional machine learning method or using American clinical data, such as the Surveillance, Epidemiology, and End Results (SEER) database [10–12].

We aimed to compare several survival algorithms based on NN and develop a deep learning survival model for colorectal cancer patients (DeepCRC) using Asian clinical data. It might offer advice for Asian doctors on patients' therapeutic decisions, to avoid unnecessary treatment and complications such as a stoma.

Methods. ***Study design and data source.*** This study was designed as a retrospective cohort study. Patients diagnosed with colorectal cancer in 2006–2014 were included and the last follow-up time was 2018. Raw clinical information was obtained from the biobank of Shanghai Outdo Biotech Company. Multivariate Imputation by Chained Equations was employed to fill in missing values (Supplementary Figure S1). All data were then divided into two cohorts randomly (Figure 1): the training cohort (80% of all) and the test cohort (20% of all). Survival models were trained using the training cohort, with validation by itself and the test cohort.

This study has been approved by the Ethics Committee (No. LW-2022-007) and individual consent for this retrospective analysis was waived.

Model training. Sex, age, size, site, grade, numbers of lymph nodes examined, numbers of positive lymph nodes, T, N, M and stage were all the clinical features included by the authors (abbreviated as ALL variables). Classical TNM variables (T, N, M and stage) were included as input features too, called TNM variables by us. Least Absolute Shrinkage and Selection Operator (LASSO) was adopted to refine variables, filtering non-zero coefficient features as LASSO variables (Age, Size, Site, Grade, Lymph nodes examined, Lymph nodes positive, T, N, M and Stage) (Supplementary Figure S2, Supplementary Table SI). Three group variables were then combined with 8 NN survival algorithms to identify the best one, with traditional Cox models conducted too as a comparison. Before building the models, categorical clinical features were recoded as dummy variables. The Adam algorithm was chosen to be an optimizer. Batch training and batch normalization were used to avoid underfitting, while dropout layers and the early stopping callback function were applied to avoid overfitting when necessary. Dropout layers could silence some neural nodes randomly and the early stopping callback function could end up training when performance did not improve during several epochs. Training curves are shown in Supplementary Figure S3.

Model evaluation. The concordance index (C-index), also known as area under the receiver operating curve (AUC), was the main criterion. The C-index close to 1.0 showed a perfect prediction, while a 0.5 or smaller one tended to randomly guess. Another indicator was the integrated Brier score, whose range was between 0 and 1, with a smaller one or near 0 representing a better performance. Each model was evaluated on the train-

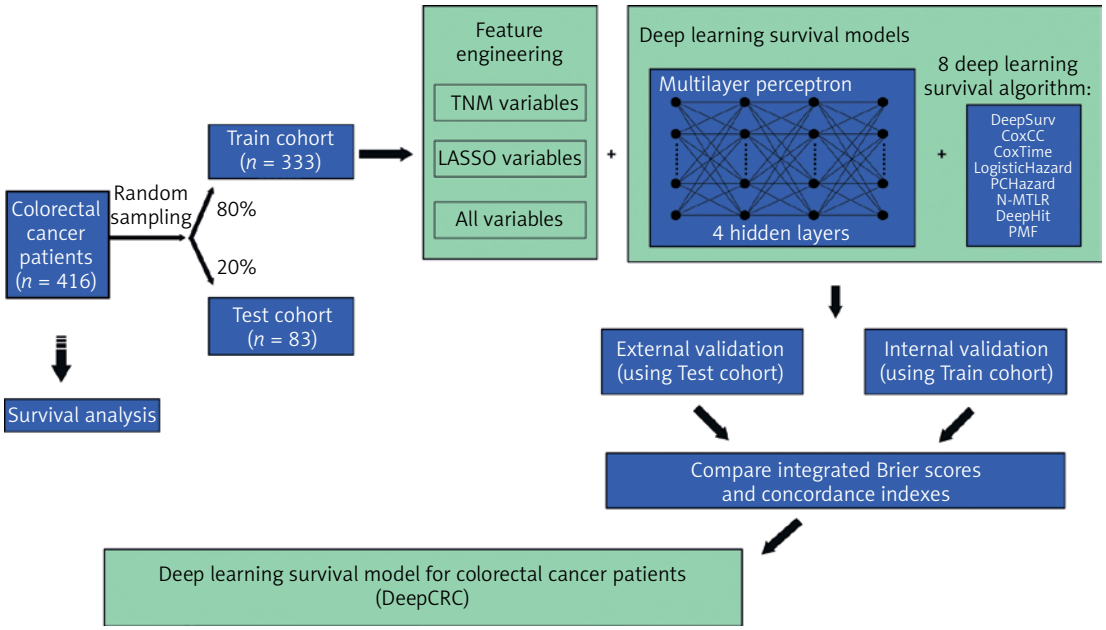

Figure 1. Schematic diagram of this study

LASSO – Least Absolute Shrinkage and Selection Operator.

ing cohort and test cohort. 1000 times bootstrap (resampling 1000 times from the training or test cohort) was taken to get precise 95% confidence intervals (CIs) of the C-index.

Data processing and statistical analysis. Missing values were visualized and imputation performed by R 4.1.2 with mice and VIM packages. LASSO regression was established with the R package glmnet. NN was constructed with python 3.9.7, pytorch and pycox. R packages (fmsb, RColorBrewer and ggplot2) were used for visualization. Two-sided $p < 0.05$ was considered statistically significant.

Results. ***Patient characteristics.*** Patients diagnosed with CRC in 2006–2014 (n = 416) were stochastically split up into two groups, the training cohort (80% of all, n = 333) and test cohort (20% of all, n = 83) (Figure 1). Table I shows the clinical characteristics of the two cohorts. The median follow-up time of the training cohort was 62 months, with that of the test cohort being 65 months. There were 156 events observed in the training cohort and 30 in the test cohort.

Model performance. As illustrated in Figure 2 and Table II, TNM variables could not reflect a patient prognosis appropriately enough even using the NN algorithm, with a C-index between 0.4756–0.6957, of which DeepSurv behaved best. When LASSO variables were inputted, the performances were boosted markedly, with the top C-in-

Table I. Demographics and clinical characteristic of two cohorts

Variable	Training cohort (n = 333) *N* (%)	Test cohort (n = 83) *N* (%)
Sex:		
Female	136 (40.84)	37 (44.58)
Male	197 (59.16)	46 (55.42)
Age:		
Median (IQR)	65 (57, 73)	66 (55.5, 75.5)
Size [mm]:		
Median (IQR)	50 (40, 70)	50 (42.5, 67.5)
Site:		
Ascending colon	53 (15.92)	9 (10.84)
Descending colon	22 (6.61)	2 (2.41)
Hepatic flexure	1 (0.3)	1 (1.2)
Ileocecal junction	7 (2.1)	2 (2.41)
Rectosigmoid junction	11 (3.3)	2 (2.41)
Rectum	186 (55.86)	53 (63.86)
Sigmoid colon	39 (11.71)	8 (9.64)
Transverse colon	4 (1.2)	1 (1.2)
Others	10 (3)	5 (6.02)
Grade:		
I	11 (3.3)	0 (0)
II	228 (68.47)	60 (72.29)
III	94 (28.23)	23 (27.71)
Lymph nodes examined:		
Median (IQR)	8 (5, 15)	7 (4, 15)
Lymph nodes positive:		
Median (IQR)	0 (0, 2)	0 (0, 1)
T:		
T1	3 (0.9)	1 (1.2)
T2	49 (14.71)	6 (7.23)
T3	181 (54.35)	57 (68.67)
T4	2 (0.6)	0 (0)
T4a	72 (21.62)	14 (16.87)
T4b	26 (7.81)	5 (6.02)
N:		
N0	189 (56.76)	51 (61.45)
N1	12 (3.6)	6 (7.23)
N1a	37 (11.11)	10 (12.05)
N1b	36 (10.81)	9 (10.84)
N1c	1 (0.3)	1 (1.2)
N2	15 (4.5)	2 (2.41)
N2a	29 (8.71)	3 (3.61)
N2b	14 (4.2)	1 (1.2)
M:		
M0	323 (97)	81 (97.59)
M1	6 (1.8)	1 (1.2)
M1a	3 (0.9)	1 (1.2)
M1b	1 (0.3)	0 (0)
Stage:		
I	44 (13.21)	7 (8.43)
II	35 (10.51)	12 (14.46)
IIA	85 (25.53)	27 (32.53)
IIB	15 (4.5)	3 (3.61)
IIC	8 (2.4)	2 (2.41)
III	26 (7.81)	7 (8.43)
IIIA	6 (1.8)	0 (0)
IIIB	77 (23.12)	21 (25.3)
IIIC	27 (8.11)	2 (2.41)
IV	6 (1.8)	1 (1.2)
IVA	3 (0.9)	1 (1.2)
IVB	1 (0.3)	0 (0)
Follow-up time:		
Median (IQR)	62 (28, 88)	65 (39.5, 90)

IQR – interquartile range.

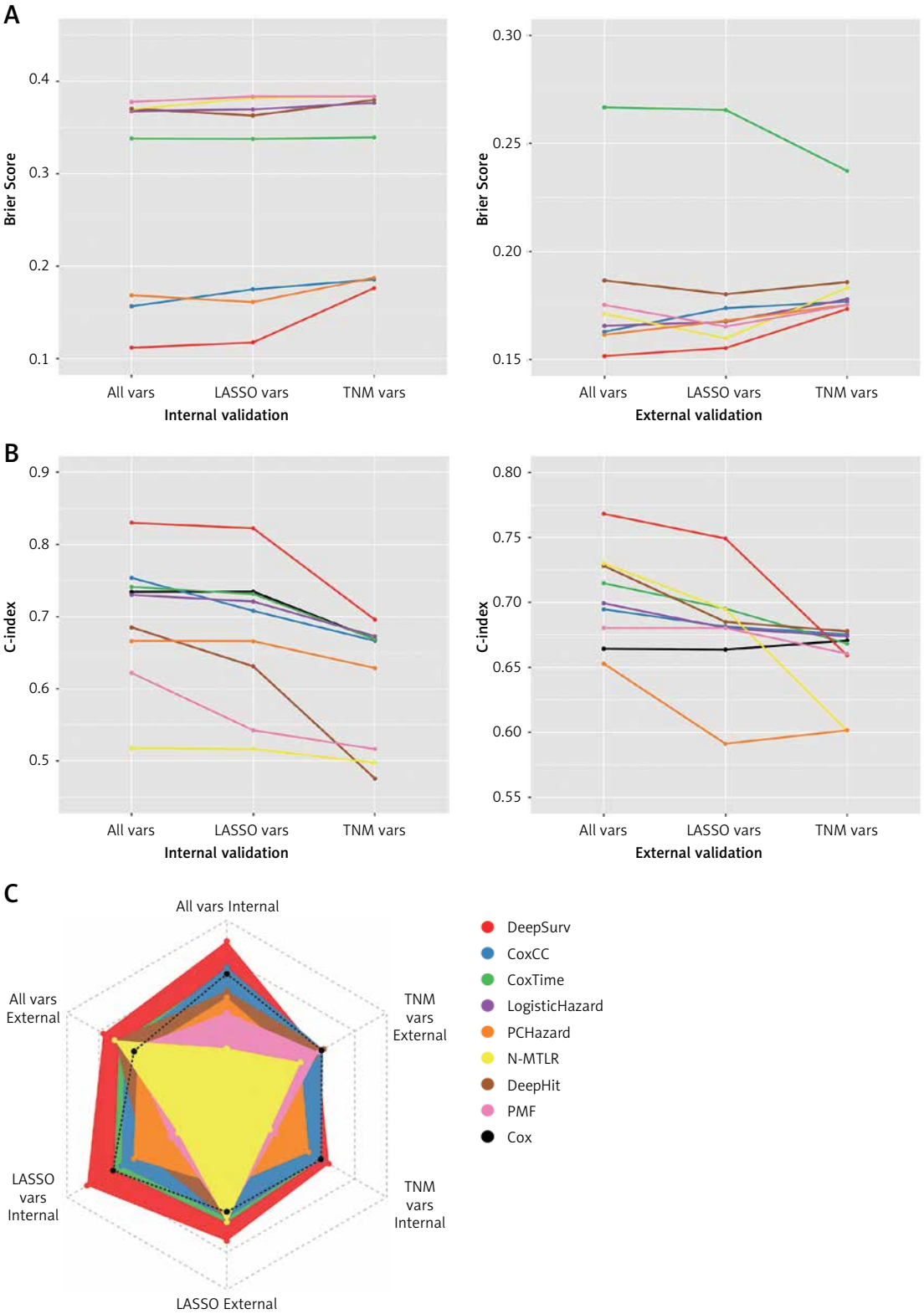

Figure 2. Performance of 8 neural network algorithms combined with 3 group variables, both internal and external validations. **A** – The Brier score of them. **B** – The concordance index of them. **C** – The radar plot showing the comparison of concordance index among these combinations

C-index – concordance index. TNM vars – T + N + M + Stage. LASSO vars – Age + Size + Site + Grade + Lymph nodes examined + Lymph nodes positive + T + N + M + Stage. All vars – Sex + Age + Size + Site + Grade + Lymph nodes examined + Lymph nodes positive + T + N + M + Stage. CoxCC – Cox Case-control Corresponding methods. PCHazard – Piecewise Constant Hazard. N-MTLR – Neural Multi-Task Logistic Regression. PMF – Probability Mass Function. LASSO – Least Absolute Shrinkage and Selection Operator.

Table II. C-index and integrated Brier score of different deep learning survival models

Models	TNM variables				LASSO variables				All variables			
	Internal validation		External validation		Internal validation		External validation		Internal validation		External validation	
	C-index	Integrated Brier Score	C-index	Integrated Brier Score	C-index	Integrated Brier Score	C-index	Integrated Brier Score	C-index	Integrated Brier Score	C-index	Integrated Brier Score
DeepSurv	0.6957	0.1763	0.6593	0.1734	0.8224	0.1174	0.7491	0.1554	0.8300	0.1118	0.7681	0.1517
CoxCC	0.6664	0.1857	0.6755	0.1768	0.7080	0.1749	0.6814	0.1738	0.7537	0.1566	0.6947	0.1629
CoxTime	0.6686	0.3392	0.6683	0.2373	0.7313	0.3375	0.6952	0.2655	0.7412	0.3379	0.7147	0.2667
LogisticHazard	0.6729	0.3764	0.6739	0.1780	0.7209	0.3695	0.6803	0.1675	0.7301	0.3675	0.6993	0.1656
PCHazard	0.6287	0.1874	0.6016	0.1754	0.6660	0.1611	0.5913	0.1680	0.6662	0.1685	0.6529	0.1615
N-MTLR	0.4975	0.3835	0.6012	0.1832	0.5164	0.3819	0.6947	0.1599	0.5179	0.3695	0.7299	0.1712
DeepHit	0.4756	0.3797	0.6778	0.1859	0.6311	0.3628	0.6850	0.1803	0.6851	0.3701	0.7281	0.1866
PMF	0.5164	0.3834	0.6606	0.1752	0.5425	0.3834	0.6803	0.1653	0.6219	0.3776	0.6803	0.1753
Cox	0.6687	–	0.6707	–	0.7347	–	0.6637	–	0.7343	-	0.6643	-

C-index, concordance index. TNM variables: T + N + M + Stage. LASSO variables: Age + Size + Site + Grade + Lymph nodes examined + Lymph nodes positive + T + N + M + Stage. All variables: Sex + Age + Size + Site + Grade + Lymph nodes examined + Lymph nodes positive + T + N + M + Stage. LASSO – Least Absolute Shrinkage and Selection Operator. CoxCC – Cox Case-control Corresponding methods. PCHazard – Piecewise Constant Hazard. N-MTLR, Neural Multi-Task Logistic Regression. PMF – Probability Mass Function.

dex up to 0.8224 in the training cohort and 0.7491 in the test cohort, from DeepSurv too. All variables were employed to conduct models finally, making some enhancement, for the C-index was determined as 0.8300 in the training cohort and 0.7681 in the test cohort by DeepSurv. Of 3 groups, ALL variables seemed to be the best indicator while DeepSurv showed the greatest potency in predicting patient OS.

After 1000 times bootstrap, DeepSurv still exhibited the best performance, with the C-index 0.8315 (95% CIs: 0.8297–0.8332) in the training cohort and 0.7719 (95% CIs: 0.7693–0.7745) in the test cohort (Supplementary Table SII).

Discussion. As a semiparametric and linear-assumption model, CPH has inherent limitations in forecasting the real word data. As the top algorithm in the machine learning field, NN has become more and more popular in the medical domain. Typical examples were application for tumor pathology or X-ray computed tomography (CT). Reasonably, researchers hoped to utilize NN to improve the accuracy of predicting cancer patients' OS. In fact, the NN survival model has shown great potential. For example, to predict urinary continence recovery after robot-assisted radical prostatectomy, Loc Trinh and colleagues compared the Cox and NN survival model DeepSurv (C-index: CPH 0.695, DeepSurv 0.708) [13]. However, there are several NN survival algorithms, but nobody has compared them yet.

Though there are already survival models for CRC, an NN model based on Asian data has not been reported but is needed. Simultaneously, we hoped to identify the best one based on our collected clinical features, by comparing 8 frequent NN survival algorithms. DeepSurv had the highest C-index in all 8 algorithms in both cohorts (0.8300 in the training cohort and 0.7681 in the test cohort). The codes we used have been uploaded to Github, hoping it will offer some help for doctors not only for CRC but also other cancers.

There were some limitations in this study. Family history, lifestyle and some biomarkers are important reasons for colorectal carcinogenesis, possibly influencing prognosis, but they were not considered in this study [14, 15]. The sample size of this study was moderate. It is better to validate DeepCRC using prospective data.

Collectively, this study pioneered the use of 8 NN survival models with real Asian data for predicting CRC patients' OS. The prediction of OS might offer a reference for doctors on treatment options.

In conclusion, we utilized and compared 8 deep learning survival models to predict CRC patients' survival (DeepCRC) using Asian data. The DeepCRC model had good performance in predicting CRC patients' overall survival.

Acknowledgments

Wei Li, Shuye Lin have equal contribution to the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021; 71: 209-49.
2. Xia C, Dong X, Li H, et al. Cancer statistics in China and United States, 2022: profiles, trends, and determinants. Chin Med J 2022; 135: 584-90.
3. Xiao WW, Zhang LN, You KY, et al. A low lymphocyte-to-monocyte ratio predicts unfavorable prognosis in pathological T3N0 rectal cancer patients following total mesorectal excision. J Cancer 2015; 6: 616-22.
4. Puppa G, Sonzogni A, Colombari R, Pelosi G. TNM staging system of colorectal carcinoma: a critical appraisal of challenging issues. Arch Pathol Lab Med 2010; 134: 837-52.
5. Gong P, Chen C, Wang Z, et al. Prognostic significance for colorectal carcinoid tumors based on the 8th edition TNM staging system. Cancer Med 2020; 9: 7979-87.
6. Shia J, Klimstra DS, Bagci P, Basturk O, Adsay NV. TNM staging of colorectal carcinoma: issues and caveats. Semin Diagn Pathol 2012; 29: 142-53.
7. Ahmed Farag AF, Elbarmelgi MY, Azim HA, Abozeid AA, Mashhour AN. TNMF versus TNM in staging of colorectal cancer. Int J Surg 2016; 27: 147-50.
8. Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC Med Res Methodol 2018; 18: 24.
9. She Y, Jin Z, Wu J, et al. Development and validation of a deep learning model for non-small cell lung cancer survival. JAMA Netw Open 2020; 3: e205842.
10. Yu H, Huang T, Feng B, Lyu J. Deep-learning model for predicting the survival of rectal adenocarcinoma patients based on a surveillance, epidemiology, and end results analysis. BMC Cancer 2022; 22: 210.
11. Mohammed M, Mboya IB, Mwambi H, Elbashir MK, Omolo B. Predictors of colorectal cancer survival using cox regression and random survival forests models based on gene expression data. PLoS One 2021; 16: e0261625.
12. Li C, Pei Q, Zhu H, et al. Survival nomograms for stage III colorectal cancer. Medicine 2018; 97: e13239.
13. Trinh L, Mingo S, Vanstrum EB, et al. Survival analysis using surgeon skill metrics and patient factors to predict urinary continence recovery after robot-assisted radical prostatectomy. Eur Urol Focus 2022; 8: 623-30.
14. Lee CH, Tseng PL, Tung HY, et al. Comparison of risk factors between colon cancer and rectum cancer in a single medical center hospital, Taiwan. Arch Med Sci 2020; 16: 102-11.
15. Zhang T, Cui G, Yao YL, et al. Value of CNRIP1 promoter methylation in colorectal cancer screening and prognosis assessment and its influence on the activity of cancer cells. Arch Med Sci 2017; 13: 1281-94.