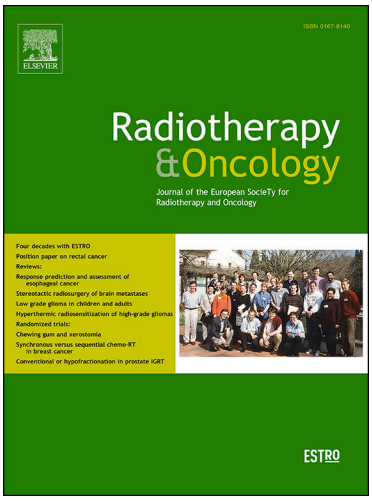
Journal Pre-proofs



Original Article

A Deep Learning Risk Prediction Model for Overall Survival in Patients with Gastric Cancer: A Multicenter Study

Liwen Zhang, Di Dong, Wenjuan Zhang, Xiaohan Hao, Mengjie Fang, Shuo Wang, Wuchao Li, Zaiyi Liu, Rongpin Wang, Junlin Zhou, Jie Tian

PII: S0167-8140(20)30333-9
DOI: https://doi.org/10.1016/j.radonc.2020.06.010
Reference: RADION 8374

To appear in: *Radiotherapy and Oncology*

Received Date: 5 January 2020
Revised Date: 8 June 2020
Accepted Date: 9 June 2020

Please cite this article as: Zhang, L., Dong, D., Zhang, W., Hao, X., Fang, M., Wang, S., Li, W., Liu, Z., Wang, R., Zhou, J., Tian, J., A Deep Learning Risk Prediction Model for Overall Survival in Patients with Gastric Cancer: A Multicenter Study, *Radiotherapy and Oncology* (2020), doi: https://doi.org/10.1016/j.radonc.2020.06.010

A Deep Learning Risk Prediction Model for Overall Survival in Patients with Gastric Cancer: A Multicenter Study

Liwen Zhang[1,2#], Di Dong[1,2#], Wenjuan Zhang[3,4#], Xiaohan Hao[2], Mengjie Fang[1,2], Shuo Wang[2,5], Wuchao Li[6], Zaiyi Liu[7*], Rongpin Wang[6*], Junlin Zhou[3,4*], Jie Tian[2,5*]

1. School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China.
2. CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.
3. Department of Radiology, Lanzhou University Second Hospital, Lanzhou 730030, China
4. Key Laboratory of Medical Imaging of Gansu Province, Lanzhou 730030, China
5. Beijing Advanced Innovation Center for Big Data-Based Precision Medicine, School of Medicine, Beihang University, Beijing 100191, China
6. Department of Radiology, Guizhou Provincial People's Hospital, Guiyang 550002, China
7. Department of Radiology, Guangdong General Hospital, Guangzhou 510080, China

# These authors contributed equally to this work.

Corresponding authors

Jie Tian, PhD

FAIMBE, FIAMBE, FIEEE, FSPIE, FOSA, FIAPR

CAS Key Laboratory of Molecular Imaging, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China.

Tel: +86-10-82618465

Fax: +86-10-82618465

Email: jie.tian@ia.ac.cn

Junlin Zhou, MD

Department of Radiology, Lanzhou University Second Hospital, Lanzhou 730030, PR China

Tel: +86-0931-8942476

Fax: +86-0931-8942476

E-mail: ery_zhoujl@lzu.edu.cn

Rongpin Wang, MD

Department of Radiology, Guizhou Provincial People's Hospital, Guiyang 550002, China

Tel: +86-0851-85273763

Fax: +86-0851-85273763

E-mail: wangrongpin@126.com

Zaiyi Liu, MD

Department of Radiology, Guangdong General Hospital, Guangzhou 510080, China

Tel: +86-20-83870125

Fax: +86-20-83870125

E-mail: zyliu@163.com

Abstract

Background and purpose: Risk prediction of overall survival (OS) is crucial for gastric cancer (GC) patients to assess the treatment programs and may guide personalized medicine. A novel deep learning (DL) model was proposed to predict the risk for OS based on computed tomography (CT) images.

Materials and methods: We retrospectively collected 640 patients from three independent centers, which were divided into a training cohort (center 1 and center 2, n=518) and an external validation cohort (center 3, n=122). We developed a DL model based on the architecture of residual convolutional neural network. We augmented the size of training dataset by image transformations to avoid overfitting. We also developed radiomics and clinical models for comparison. The performance of the three models were comprehensively assessed.

Results: Totally 518 patients were prepared by data augmentation and fed into DL model. The trained DL model significantly classified patients into high-risk and low-risk groups in training cohort (*P*-value<0.001, concordance index (C-index): 0.82, hazard ratio (HR): 9.79) and external validation cohort (*P*-value<0.001, C-index:0.78, HR: 11.76). Radiomics model was developed with selected 24 features and clinical model was developed with three significant clinical variables (*P*-value<0.05). The comparison illustrated DL model had the best performance for risk prediction of OS according to the C-index (training: DL vs Clinical vs Radiomics=0.82 vs 0.73 vs 0.66; external validation: 0.78 vs 0.71 vs 0.72).

Conclusion: The DL model is a powerful model for risk assessment, and potentially serves as an individualized recommender for decision-making in GC patients.

Keywords: Gastric cancer; Deep learning; Overall survival; Individualized treatment; Computed tomography

Introduction

Gastric cancer (GC) is one of the most common gastrointestinal malignancies worldwide. Although its incidence has decreased, GC still serves as the third leading cause of cancer-associated deaths, particularly in Eastern Asia [1]. Currently, patients with advanced GC are recommended to receive surgical resection, adjuvant chemotherapy and radiotherapy for improvement of the treatment in line with the US National Comprehensive Cancer Network guidelines [2]. However, previous studies revealed that the rates of 5-year survival are still poor, and surgical morbidity is high [3], which have led to wide investigation for survival analysis.

The state-of-the-science tumor-node-metastasis (TNM) staging system (8^{th} edition) of GC promulgated by The American Joint Committee on Cancer (AJCC) is widely used as the gold standard for prognostic evaluation and survival risk stratification [4]. However, the authors indicated that the manual is not an exact science, which is the ongoing work and will be updated to reflect the state-of-the-art changing [4]. Particularly, for personalized medicine, The AJCC Personalized Medicine Core (PMC) committee has been increasingly conscious of the necessity for more individualized predictions than those presented by ordinal cancer staging systems based on risk models constructed by machine learning approach [5]. Overall survival (OS) was required by The AJCC PMC as the outcome being predicted for the risk models.

Recently, related works have focused on non-invasive methods of imaging, especially for computed tomography (CT), which is the routinely used modality for staging and risk assessment [6]. Radiomics, an emerging field, is an accepted method to analyze the medical images by extracting amounts of quantitative features [7]. Previous study found that radiomic features extracted from CT images had prognostic value for overall survival in patients with lung cancer [8]. Zhang et al. developed a radiomic nomogram to predict early recurrence in GC patients following curative

resection [9]. Some studies have demonstrated that radiomics method was a meaningful tool associated with tumor prognosis in patients with nasopharyngeal cancer and gastric cancer [10, 11]. Nevertheless, standard procedures of radiomics method need accurate delineation for segmentation and retest the stability for features, which may cause variability and inconsistent reproducibility [12]. Thus, it is relatively convenient and labor-saving to develop a tool for survival analysis, which can lower the delineation standard and tailor training process to train the model by feeding segmented CT images into the network and extracting the features by itself [13].

Currently, studies on medical image analysis are undergoing a transformation from engineering of feature extraction to self-learning. In particular, deep learning (DL), a state-of-the-art methodology, has attracted much attention and achieved huge breakthroughs in a wide range of computer vision task and clinical applications [14]. Bello et al. demonstrated DL method can be applied to develop a motion model to efficiently predict survival [15]. DL method has been also used for the screening the GC patient focusing on endoscopic image-based analysis [16]. However, the implement of DL method for the risk prediction of OS in GC patients based on CT images remains unclear.

In this study, we developed a DL model for risk prediction of OS based on the widely recognized residual convolutional neural network (CNN) [17]. We also constructed a potential individualized recommender system to provide recommendations for decision-making.

Materials and Methods

Figure 1 shows an overview of this study via the DL model in combination with the TNM staging system for the individualized treatment.

Patients

Ethical approval was respectively received for the Institutional Review Board of each center, and informed consent from patients was waived. This was a retrospective

multicenter study. A total of 640 consecutive patients who were pathologically diagnosed with GC from June 2010 to April 2019 were enrolled from three independent centers. We divided eligible patients into a training cohort (n=518, from center 1 and center 2) and an external validation cohort (n=122 from center 3), which is shown in supplementary Figure A1. Supplementary Table A1 and Part 2 show the parameters for CT images and the details of follow-up for OS. Characteristics in the training and external validation cohorts are shown in the Table A2.

Image segmentation

We used the software ITK-SNAP for segmentation [18]. For each patient, we selected a slice of CT image with largest tumor region and nearest upper and lower slices in portal venous phase by two experienced radiologists and outlined them with three rough rectangle boxes. In order to avoid coarse label for each patch, the region of interest was acquired at first. Afterwards, the input image for the deep learning model was the region of interest. For constructing the radiomics model, we manually delineated precisely the tumor region of the slice with largest tumor region again. The diagram of segmentation is shown in Figure 2.

Model construction

We constructed a DL model based on 18-layers residual CNN with the input of segmented CT images (size: 224*224) [17]. The model consisted of 8 residual blocks, which have the “short cut” for transmitting gradient efficiently and accelerating the convergence of the network (Figure 2). We tailored the dense and dropout layers at the top of the model. We also defined the specialized loss function (Supplementary Formula A1) to train the model for risk prediction. Some techniques including data augmentation and fine-tuning were used to train the model and avoid overfitting. More details regarding the training procedure can be found in the Supplementary Part 4. For comparison, we also constructed the radiomics model (Figure 2B) based on hand-crafted features and clinical models for comparison. The output of each model，named risk score for each GC patient, represented the hazard degree for occurrence of

the endpoint of interest.

Assessment of prognostic performance for DL model

To investigate the potential association between the proposed DL model and OS, we depicted Kaplan-Meier (KM) curves. For each patient, the cutoff of median risk score was obtained in the training cohort. Patients with the scores lower than the cutoff were classified into the low-risk group, while others were classified into the high-risk group.

Furthermore, we performed stratification analysis to validate the performance of the DL model in different subgroups (T stage, N stage, TNM stage, and adjuvant chemotherapy). We employed visualization techniques to present the self-learned feature maps inside the DL model [19]. We developed a risk score grading tool based on a widely used nomogram [20]. To show the network benefit, the clinically accepted tool of decision curve analysis (DCA) was applied to verify the prognostic value of the DL model [20]. We calculated the Harrell's concordance index (C-index) and hazard ratio (HR) to evaluate the performance of the three models. Finally, we proposed an individualized recommender system for potential clinical application.

Statistical analysis

We performed the statistical analysis with R software (http://www.R-project.org). The features and clinical variables were compared using the Mann-Whitney U test. KM curves were compared by Log-Rank test. Moreover, the G-rho rank test was used for calculation of the HR [21]. We also compared the C-index of the DL model with other models by a non-parametric test. The result was considered statistically significant when the *P*-value (from two-sided tests) was less than 0.05.

Results

Schoenfeld residuals test demonstrated that each clinical variable was eligible to use Cox regression for univariable and multivariable analysis (Figure A2). T stage, N

stage, and adjuvant chemotherapy were significant (*P*-value<0.05, Table A3) for construction of clinical model. In training cohort, median survival time was 28 months. In external validation cohort, the median survival time was 56 months.

The DL model with residual blocks and the identity mapping was constructed to show the learning capability of risk prediction. A total of 12432 images were generated by data augmentation and fed into DL model. Finally, the risk score was output for each patient. Further details regarding the development of the radiomics model are shown in Supplementary Part 7.

The DL model could classify all patients into two different risk subgroups (the low-risk and high-risk) in the training cohort (*P*-value<0.001, C-index: 0.82, 95% confidence interval (CI) 0.80-0.84, HR: 9.79, 95%CI 7.15-13.41) and external validation cohort (*P*-value<0.001, C-index: 0.78, 95%CI 0.72-0.83, HR: 11.76, 95%CI 4.23-32.71). KM curves for DL model are shown in Figure 3. Clinical data analysis of different risk groups in the both cohorts are shown in Table 1. In the training cohort, median survival time was 14 months in the high-risk group, and 45 months in low-risk group. In the external validation cohort, the median survival time in the high-risk group was 45 months, and 66 months in low-risk group. The stratification analysis revealed that the DL model also had good performance for risk prediction in different subgroups pertaining to N stages (Figures A3), T stages (Figures A4), TNM stages (Figure A5), and adjuvant chemotherapy (Figure A6).

DL model could learn discriminative features for GC patients in different risk groups with different survival time. Examples for two patients were shown in the Figure 4, one from the high-risk group and one from the low-risk group in line with the risk scores predicted by the three models. The feature maps extracted from shallow to deep layers of the DL model were visualized. Highly responsive areas colored red of a region of interest (ROI) were found the different in two risk groups. With the number of layers of the network increasing, the DL model can focus on the highlights in the ROI with small sized feature maps by convolution and pooling. The

high-risk groups with corresponding risk scores (DL vs Clinical vs Radiomics: 0.98 vs 0.6 vs 0.62) and low-risk groups with corresponding risk scores (DL vs Clinical vs Radiomics: 0.1 vs 0.28 vs 0.49) were obtained by the three models.

Clinical data analysis in Table A2 shows that the HR for adjuvant chemotherapy was 0.62 (95% CI: 0.46-0.82), which revealed the adjuvant chemotherapy was a good prognostic factor for GC patients. DL model also showed consistent risk prediction by further stratification. According to the analysis of multiple factors of N stage and adjuvant chemotherapy, the further results (Figure A7) shows that GC patients were also divided into high and low cumulative hazard subgroups by the DL model. We also found that the DL model can performed well according to the T stage and adjuvant chemotherapy (Figure A8). In each subgroup, the findings showed that GC patients in low-risk groups with lower cumulative hazard grouped by DL model had better OS than the high-risk.

To evaluate the performance of the three models, the comparison for KM curves is shown in Figure 3. The cut-off obtained in the training cohort was 0.668, 0.502 and 0.504 for the deep learning model, clinical model and radiomics model respectively. We depicted the distribution of risk scores for patients predicted by three models (Figure 4A). In the distribution shape of the risk score for the DL model in the training cohort, all patients were divided into two subgroups, wherein patients in high-risk group were centralized. Moreover, in Figure 4B and 4C, the DL model shows the best capability for prediction with the highest C-index in the training cohort (DL vs Clinical vs Radiomics: 0.82 vs 0.73 vs 0.66), wherein the comparison of C-index was significant (P-value<0.01). Meanwhile, C-index of the DL model also outperformed the others in the external validation cohort (DL vs Clinical vs Radiomics: 0.78 vs 0.71 vs 0.72) with a significant difference between the DL model and the clinical model (P-value<0.05).

As is shown in Figure 4C, the DL model had the highest HR both in training cohort: (9.79 vs 3.84 vs 2.48) and external validation cohort (11.76 vs 3.57 vs 5.86),

which indicated the high-risk groups predicted by DL model had higher hazard of death than the high-risk groups predicted by other models. Furthermore, in comparisons of the time-dependent receiver operating characteristic (ROC) curves (1-year, 2-year and 3-year) for three models, we found that the performance of the DL model equally outperformed the other models in both cohorts (Figure A9).

The nomogram, calibrations and DCA curves of DL model were depicted in Figure 5, which shows good performance for risk prediction. The DCA indicated that the DL model provided a greater net benefit than other models for the patients. Therefore, we constructed the individualized grading rule of nomogram to divide GC patients into low-risk and high-risk subgroups. After all GC patients were divided into two groups by the individualized grading rule, we constructed a deep learning-aided recommender by calculating the difference ($D_\beta(x)$ - D_{risk}) to measure the degree of risk in the subgroups (Supplementary Figure A10).

In order to show the generalizability of the model, we constructed deep learning (DL) model again, where we combined center 1 and 3 as the training cohort (n = 459) and used center 2 as external validation cohort (n = 181). As shown in the follow figure, the DL model significantly classified patients into high-risk and low-risk groups in training cohort (Supplementary Figure A11, *P*-value<0.001, concordance index (C-index): 0.77, 95% confidence interval (CI):0.74-0.79, hazard ratio (HR): 5.54 , 95%CI: 4.07-7.54) and external validation cohort (*P*-value<0.001, C-index: 0.76, 95%CI 0.71-0.80, HR: 6.91, 95%CI 4.06-11.78).

Discussion

In this study, we investigated the performance of a DL model using CT images, with the aim of improving the prediction of OS for GC patients. The DL model showed encouraging outcomes with regard to its capability to stratify GC patients into two groups with discrepant prognosis in the training and external validation cohorts compared with other models. We found that high-risk groups had poor OS, whereas

low-risk groups better. To further visualize and interpret the dynamic change inside the DL model, feature maps were vividly visualized and represented. For the standard treatment of adjuvant chemotherapy, covariate analysis for the DL model shows potential guidelines for GC patients.

We proposed a DL model based on residual network, which was demonstrated that it was suitable to predict the risk for GC patients. We implemented several methods to train the model (Supplementary Part 4). He et al. have demonstrated that residual block and the identity mapping can improve the learning capability, and address the degradation problem [17]. Our outcomes revealed that residual network, in some cases, could also address the degradation problem for CT images analysis, and data augmentation was useful for enlarging the training data to cope with the problem of overfitting. Meanwhile, the techniques of dropout and fine-tuning were also efficient to improve the robustness for the DL model based on limited CT images.

Although the golden standard for treatment of GC patients is AJCC TNM staging system,[4] the AJCC has realized that risk model for OS is necessary for more individualized probabilistic prognostication [5]. In particular, for personalized treatment, previous studies have implicated that the TNM staging system have some drawbacks [22]. For instance, although the patients belonged to confirmed subgroup (T stage=T2, N stage =N0, TNM stage =IB), we can't obtain the further information about degree of risk for each patient or the different risk groups they belonged to, which may lead to suboptimal recommendations for individualized treatment. Conversely, the DL model can classify the patients into different risk groups and the recommender provides recommendation for individualized treatment combined with TNM staging system. According to the stratification analysis with multiple factors of the N stage, T stage, and adjuvant chemotherapy, the findings revealed that the DL model is a powerful predictor for risk prediction, which have the potential to serve as a model-based reference index for an updating TNM staging system to improve

clinical decision making.

Currently, the popular method of radiomics plays an important role in prognostic analysis [23]. However, elaborate delineation by radiologists of the ROI hinders deployment of segmentation in clinical practice. In practice, the work of segmentation for our proposed model is easy and time-saving to complete, since we do not require the tumor to be precisely delineated. Hence, the DL model is considered relatively as an easy-to-use and labor-saving tool for clinical application. In addition, compared with hand-crafted radiomics features, the feature maps were learned automatically from the shallow to deep convolutional layers by the DL model, including simple low-level features to complex high-level features. Hence, our study presents a promising approach.

In addition, Cox Proportional Hazard (CPH) Model are widely used for survival analysis [24]. However, the assumption of CPH model that logarithmic HR is linearly correlated with each risk factor is restrictive. While universally-applicable methods, such as DL method, can construct robust model without any assumption. Katzman et al. illustrated that DL model (DeepSurv) showed good performance and can provide personalized recommendations based on simulated and real survival data [25]. Yousefi et al. reported that DL model showed good performance to learn information from diseases for survival analysis with public molecular data [26]. Kim et al. applied the DeepSurv for survival analysis in oral squamous cell carcinoma (SCC) patients, and the model outperformed the random survival forest (RSF) and the CPH models [27]. Matsuo et al. investigated 40 clinical features and indicated that DL model can be a potential tool for survival prediction in women with cervical cancer, which showed superior performance than CPH model [28]. Sun et al. proposed a multimodal DL model by integrating Multi-dimensional Data (clinical and gene) for the prognosis prediction of breast cancer, and achieved a better performance than other existing methods with single-dimensional data [29]. Nie et al. constructed a 3D DL model for feature extraction with multi-modality brain images of glioma patients,

and experimentally found that the DL method could learned discriminative features from multi-modality images for accurate prediction of the OS time [30]. Previous study exploited a multi-channel 3D CNN model to extract self-learned features from multi-modal brain images [31]. They found multi-channel deep survival network is powerful for prediction of OS time. Yao et al. proposed a deep correlational survival model to handle multi-modality data effectively, and the result demonstrated that the learned interactions can affect survival outcomes [32]. We also demonstrated that DL model outperformed the model constructed based on hand-crafted features and the clinical model based on clinical risk factors.

Despite the encouraging performance of the DL model, there are several limitations. Our model was developed based on the patients of Asian race, and further validation across other races should be studied. Although the segmentation was time-saving for the point of the comparison, the work has defects. The further work should be done for comparison to show the advantage of deep learning method for survival model. Meanwhile, the DL model was demonstrated here with only a limited dataset, and a larger dataset should be collected to validate a more robust performance. Additionally, the model was constructed only based on preoperative CT images, which may show more significant findings in combinations with pathological image and other types. We only employed three slices in each patient to construct model. Although three-dimensional delineation is time-consuming, the further work should be explored. Above all, although the work of interpretation and visualization was shown, the more acceptable and friendly approach for interpretability should be investigated.

In conclusion, the DL model can provide CT-based prognostic risk scores related to the OS of GC patients, and the findings demonstrated higher prognostic value than clinical and radiomics models. Most notably, our individualized recommender based on the DL model was validated through diverse verification, wherein it showed powerful prognostic ability. Therefore, the recommender is a potential tool to assist

clinicians with therapeutic decision-making and individualized treatment.

Funding Sources

This work was supported by the National Key R&D Program of China (2017YFC1308700, 2017YFA0205200, 2017YFC1309100, 2017YFA0700401), National Natural Science Foundation of China (81971776, 91959130, 81771924, 81925023, 81227901, 81772006, 81771912, 81930053, 81960314), the Beijing Natural Science Foundation (L182061), the Bureau of International Cooperation of Chinese Academy of Sciences (173211KYSB20160053), the Instrument Developing Project of the Chinese Academy of Sciences (YZ201502), and the Youth Innovation Promotion Association CAS (2017175), Technology Foundation of Guizhou Province (QKHJC[2016]1096).

Declaration of Interests

The authors declare no potential conflicts of interest.

References

[1] Bray F, Ferlay J, Soerjomataram I, et al. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians. 2018;68(6):394-424.

[2] Ajani JA, D'Amico TA, Almhanna K, et al. Gastric cancer, version 3.2016, NCCN clinical practice guidelines in oncology. Journal of the National Comprehensive Cancer Network. 2016;14(10):1286-312.

[3] Tegels JJ, De Maat MF, Hulsewé KW, et al. Improving the outcomes in gastric cancer surgery. World Journal of Gastroenterology: WJG. 2014;20(38):13692.

[4] Amin MB, Greene FL, Edge SB, et al. The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA: a cancer journal for clinicians. 2017;67(2):93-9.

[5] Kattan MW, Hess KR, Amin MB, et al. American Joint Committee on Cancer acceptance criteria for inclusion of risk models for individualized prognosis in the practice of precision medicine. CA: a cancer journal for clinicians. 2016;66(5):370-4.

[6] Smyth E, Verheij M, Allum W, et al. Gastric cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. Annals of oncology. 2016;27(suppl_5):v38-v49.

[7] Lambin P, Rios-Velazquez E, Leijenaar R, et al. Radiomics: extracting more information from medical images using advanced feature analysis. European journal of cancer. 2012;48(4):441-6.

[8] van Timmeren JE, Leijenaar RT, van Elmpt W, et al. Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images. Radiotherapy and Oncology. 2017;123(3):363-9.

[9] Zhang W, Fang M, Dong D, et al. Development and validation of a CT-based radiomic nomogram for preoperative prediction of early recurrence in advanced gastric cancer. Radiotherapy and Oncology. 2020;145:13-20.

[10] Dong D, Zhang F, Zhong L-Z, et al. Development and validation of a novel MR imaging predictor of response to induction chemotherapy in locoregionally advanced nasopharyngeal cancer: a randomized controlled trial substudy (NCT01245959). BMC medicine. 2019;17(1):190.

[11] Jiang Y, Chen C, Xie J, et al. Radiomics signature of computed tomography imaging for prediction of survival and chemotherapeutic benefits in gastric cancer. EBioMedicine. 2018;36:171-82.

[12] Lambin P, Leijenaar RT, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nature reviews Clinical oncology. 2017;14(12):749.

[13] Beig N, Khorrami M, Alilou M, et al. Perinodular and intranodular radiomic features on lung CT images distinguish adenocarcinomas from granulomas. Radiology. 2018;290(3):783-92.

[14] Bi WL, Hosny A, Schabath MB, et al. Artificial intelligence in cancer imaging: clinical challenges and applications. CA: a cancer journal for clinicians. 2019;69(2):127-57.

[15] Bello GA, Dawes TJ, Duan J, et al. Deep-learning cardiac motion analysis for human survival prediction. Nature machine intelligence. 2019;1(2):95.

[16] Hirasawa T, Aoyama K, Tanimoto T, et al. Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images. Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association. 2018;21(4):653-60.

[17] He K, Zhang X, Ren S, et al., editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

[18] Yushkevich PA, Piven J, Hazlett HC, et al. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage. 2006;31(3):1116-28.

[19] Yosinski J, Clune J, Nguyen A, et al. Understanding neural networks through deep visualization. arXiv preprint arXiv:150606579. 2015.

[20] Huang Y-q, Liang C-h, He L, et al. Development and validation of a radiomics nomogram for preoperative prediction of lymph node metastasis in colorectal cancer. Journal of Clinical Oncology. 2016;34(18):2157-64.

[21] Hernán MA. The hazards of hazard ratios. Epidemiology (Cambridge, Mass). 2010;21(1):13.

[22] Li W, Zhang L, Tian C, et al. Prognostic value of computed tomography radiomics features in patients with gastric cancer following curative resection. European radiology. 2019;29(6):3079-89.

[23] Dong D, Tang L, Li Z-Y, et al. Development and validation of an individualized nomogram to identify occult peritoneal metastasis in patients with advanced gastric cancer. Annals of Oncology. 2019;30(3):431-8.

[24] Cox DR. Analysis of survival data: Chapman and Hall/CRC; 2018.

[25] Katzman JL, Shaham U, Cloninger A, et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. BMC medical research methodology. 2018;18(1):24.

[26] Yousefi S, Amrollahi F, Amgad M, et al. Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. Scientific reports. 2017;7(1):11707.

[27] Kim DW, Lee S, Kwon S, et al. Deep learning-based survival prediction of oral cancer patients. Scientific reports. 2019;9(1):1-10.

[28] Matsuo K, Purushotham S, Jiang B, et al. Survival outcome prediction in cervical cancer: Cox models vs deep-learning model. American journal of obstetrics and gynecology. 2019;220(4):381. e1-. e14.

[29] Sun D, Wang M, Li A. A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data. IEEE/ACM transactions on computational biology and bioinformatics. 2018;16(3):841-50.

[30] Nie D, Zhang H, Adeli E, et al., editors. 3D deep learning for multi-modal imaging-guided survival time prediction of brain tumor patients. International conference on medical image computing and computer-assisted intervention; 2016: Springer.

[31] Nie D, Lu J, Zhang H, et al. Multi-channel 3D deep feature learning for survival time prediction of brain tumor patients using multi-modal neuroimages. Scientific reports. 2019;9(1):1-14.

[32] Yao J, Zhu X, Zhu F, et al., editors. Deep correlational learning for survival prediction from multi-modality data. International Conference on Medical Image Computing and

Computer-Assisted Intervention; 2017: Springer.

Figure legends

Figure 1 The overview of the study design.

Figure 2 Architectures of deep learning (DL) model based on residual convolutional neural network and radiomics model based on hand-crafted features.

Figure 3 Comparison of Kaplan-Meier (KM) curves for models. (A), (D): deep learning (DL) model; (B), (E): clinical model; (C), (F): radiomics model. Each vertical tick on the bottom of the KM curves represents a patient who was censored at that time.

Figure 4 Model analysis with measurable indicators. (A) Risk score distribution for the origin output of three models; (B) Comparison of the C-index between deep learning (DL) models and other models by *P*-value. The Student's t-test by R package ("survcomp", version:1.34.0) used for the comparison of the concordance indices; (C) Comparison of model performance by concordance index (C-index) and hazard ratio (HR). (D) Model interpretation and visualization for the potential association between feature maps with pathological staging information. † The HR was calculated by comparing the high-risk group with low-risk group. ‡ The clinical model was constructed based on AJCC 8th staging system in combination with the risk factors of adjuvant chemotherapy.

Figure 5 Clinical application and further validation for deep learning (DL) model. (A) Individualized grading rules for risk score based on nomogram. (B) Calibration curves of the DL nomogram in the training cohort. (C) Comparison of decision curve analysis for the DL model.

Declaration of Interests

The authors declare no potential conflicts of interest.

Highlights

1. Deep learning model is a potential tool for risk prediction.

2. Both radiomics model and deep learning model had prognostic values.

3. The deep learning model can classify the patients into low- and high-risk groups.

4. Individualized recommender is a potential tool to assist clinicians.

Table 1. Characteristics analysis by deep learning model

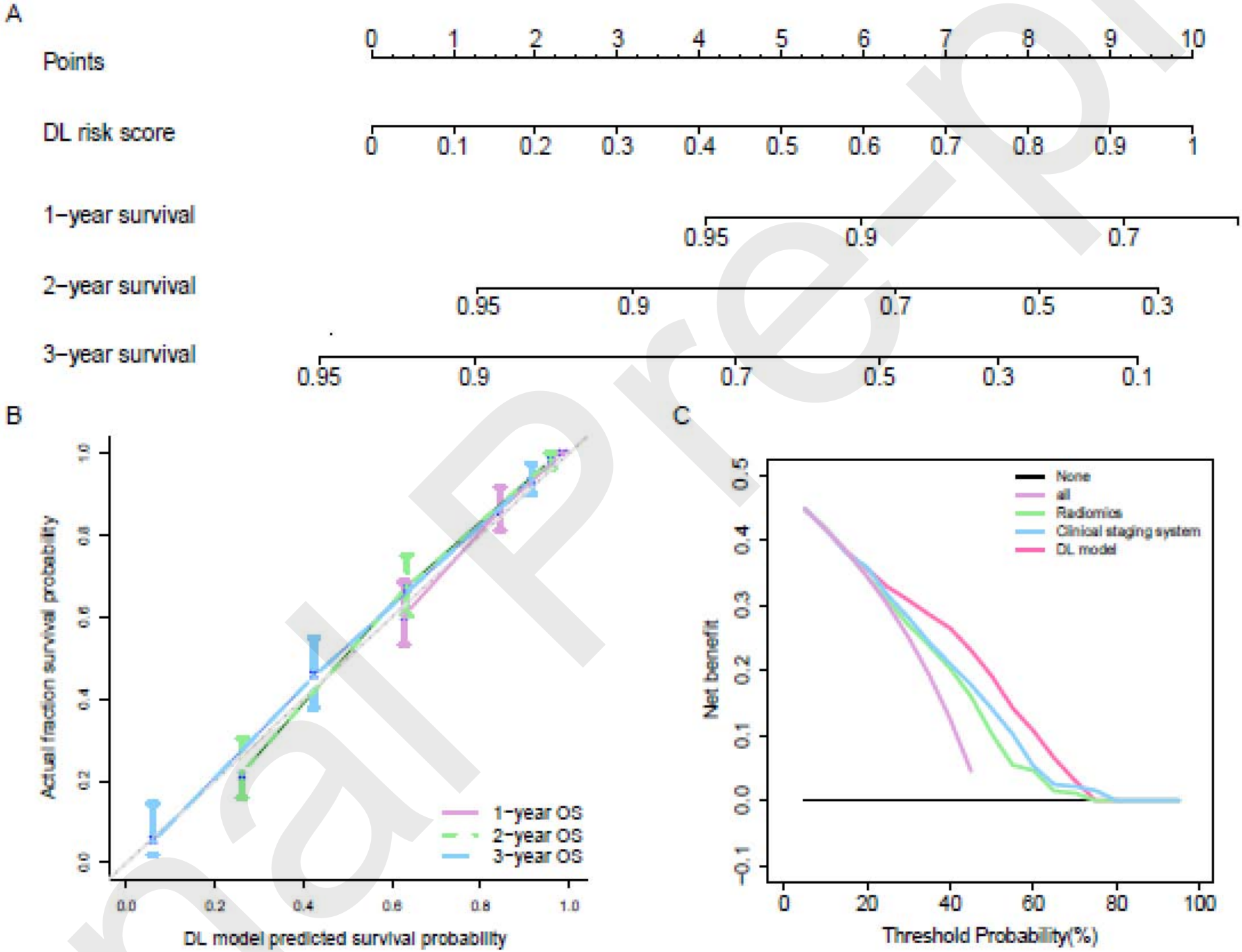
Characteristic	Training cohort		*P*-value‡	External validation cohort		*P*-value‡
	Low risk	High risk		Low risk	High risk	
Age (years)†	57(56±10)	59(56±11)	0.14	59(59±10)	58(58±12)	0.77
Gender			0.86			0.74
Male	189(73)	188(73)		32(70)	52(68)	
Female	70(27)	71(27)		14(30)	24(32)	
Tumor localization			0.49			<0.01
Proximal	46(18)	54(21)		13(28)	33(43)	
Middle	68(26)	66(25)		5(11)	13(17)	
Distal	145(56)	139(54)		28(61)	30(40)	
Tumor size			0.02			<0.01
< 5cm	173(67)	116(45)		28(61)	23(30)	
≧5cm	86(33)	143(55)		18(39)	51(67)	
NA				0(0)	2(3)	
Lymphovascular invasion			0.01			0.28
Negative	82(32)	49(19)		20(43)	31(41)	
Positive	177(68)	210(81)		26(57)	45(59)	
Differentiation			0.05			0.13
Well + moderate	127(49)	89(34)		17(37)	25(33)	
Poor + undifferentiated	132(51)	170(66)		29(63)	51(67)	
T stage			<0.01			<0.01
T1a-T1b	22(8)	1(0)		3(6)	2(2.50)	
T2	48(19)	12(5)		7(15)	2(2.50)	
T3	113(44)	88(34)		32(70)	57(75)	
T4a	76(29)	158(61)		4(9)	15(20)	
N stage			<0.01			<0.01
N0	108(42)	35(14)		10(22)	14(19)	
N1	58(22)	29(11)		15(33)	10(13)	
N2	39(15)	67(26)		9(19)	20(26)	
N3a-N3b	54(22)	128(49)		12(26)	32(42)	
TNM stage			<0.01			<0.01
I	40(15)	6(2)		7(15)	2(3)	
II	82(32)	29(11)		19(41)	23(30)	
III	137(53)	224(87)		20(44)	51(67)	
Adjuvant chemotherapy			<0.01			0.63

No	47(18)	61(24)	25(54)	37(49)
Yes	212(82)	198(76)	21(46)	39(51)
Follow-up (Month)†	45(43±14)	14(17±11)	66(67±21)	45(45±27)

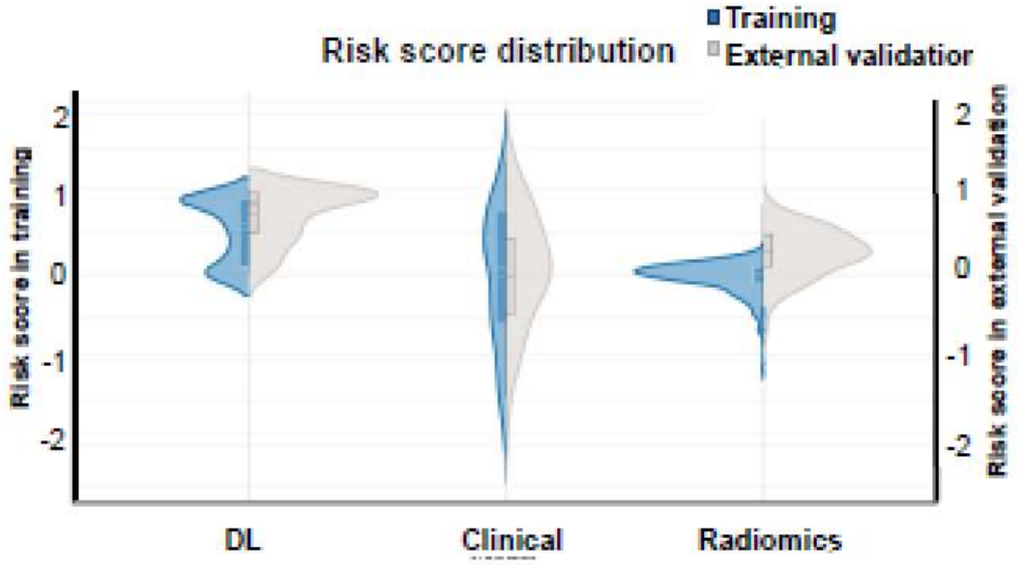
Values in parentheses are percentages (%)

† Values are median(mean+std)

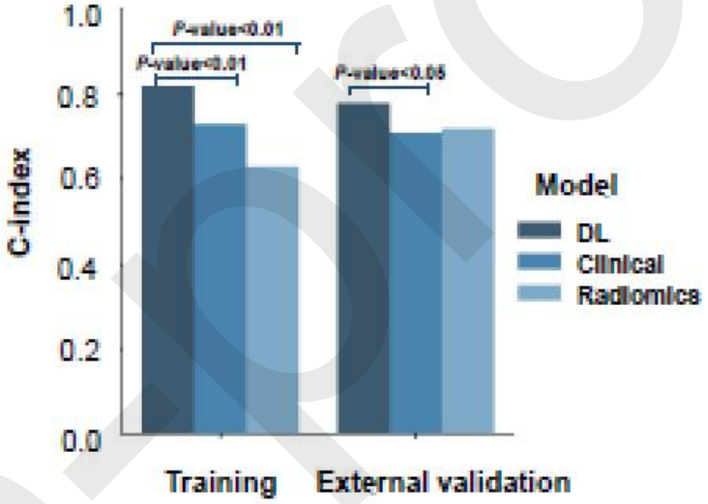
‡ Continuous variables were tested by Mann-Whitney U test and discrete variables were tested by Pearson's Chi-squared test.


A
Points
0 1 2 3 4 5 6 7 8 9 10
DL risk score
0 0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8 0.9 1
1−year survival
0.95 0.9 0.7
2−year survival
0.95 0.9 0.7 0.5 0.3
3−year survival
0.95 0.9 0.7 0.5 0.3 0.1
B
Actual fraction survival probability
DL model predicted survival probability
1−year OS
2−year OS
3−year OS
C
Net benefit
Threshold Probability(%)
None
all
Radiomics
Clinical staging system
DL model

A



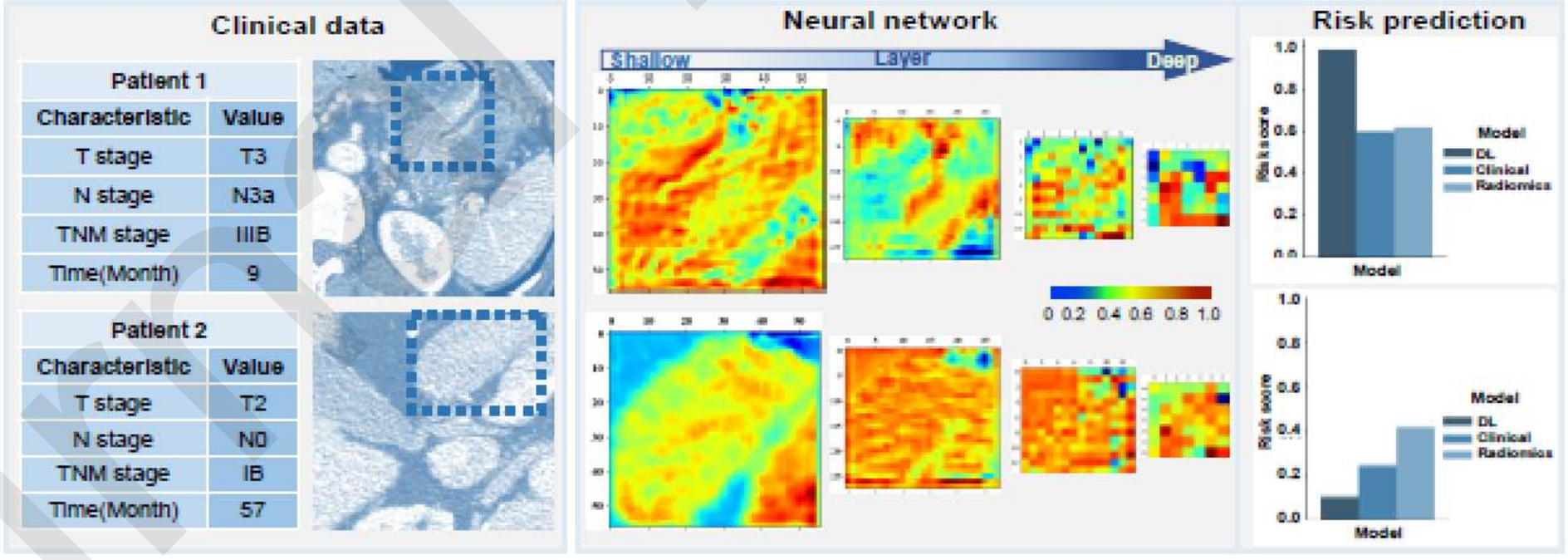
B

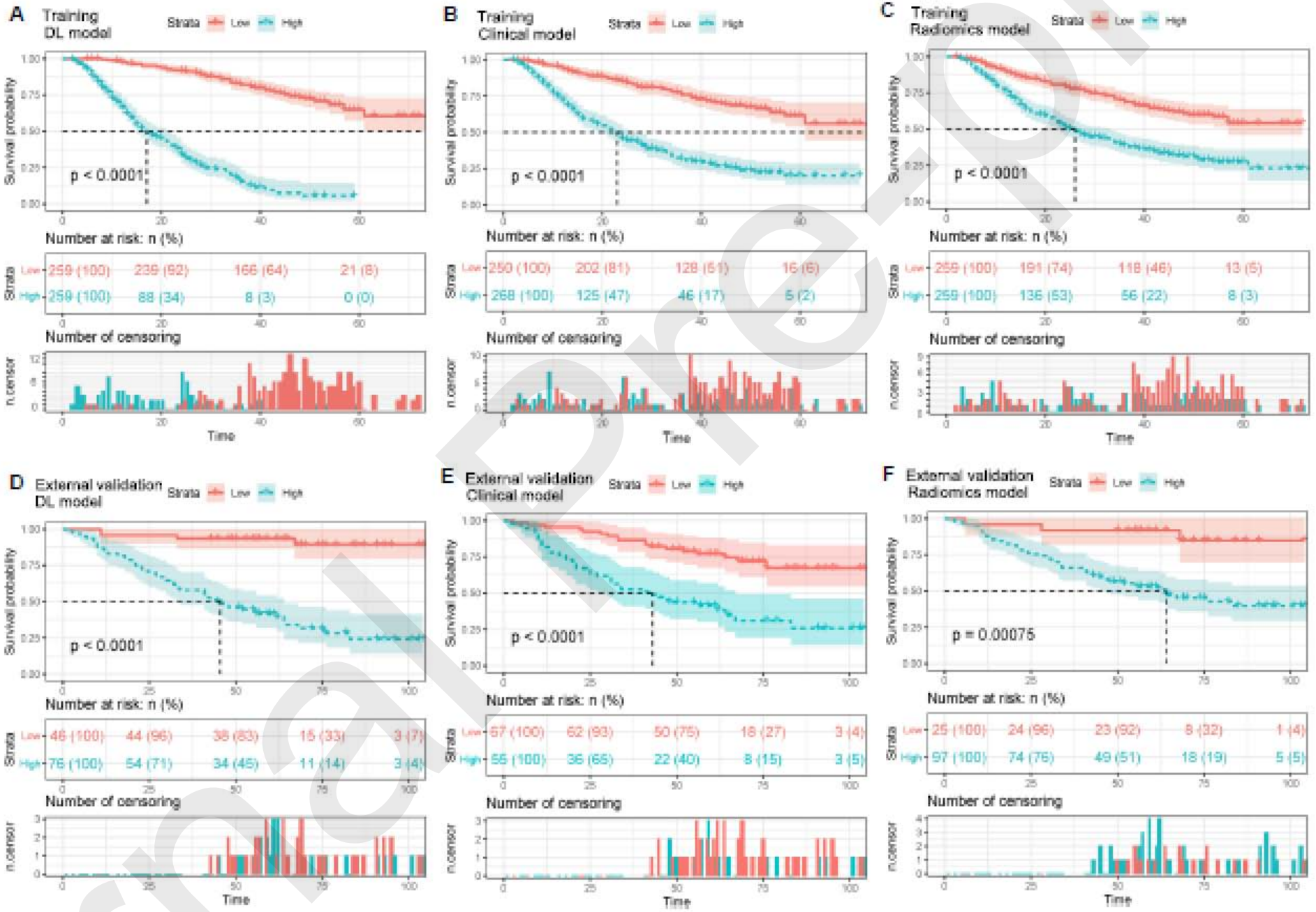


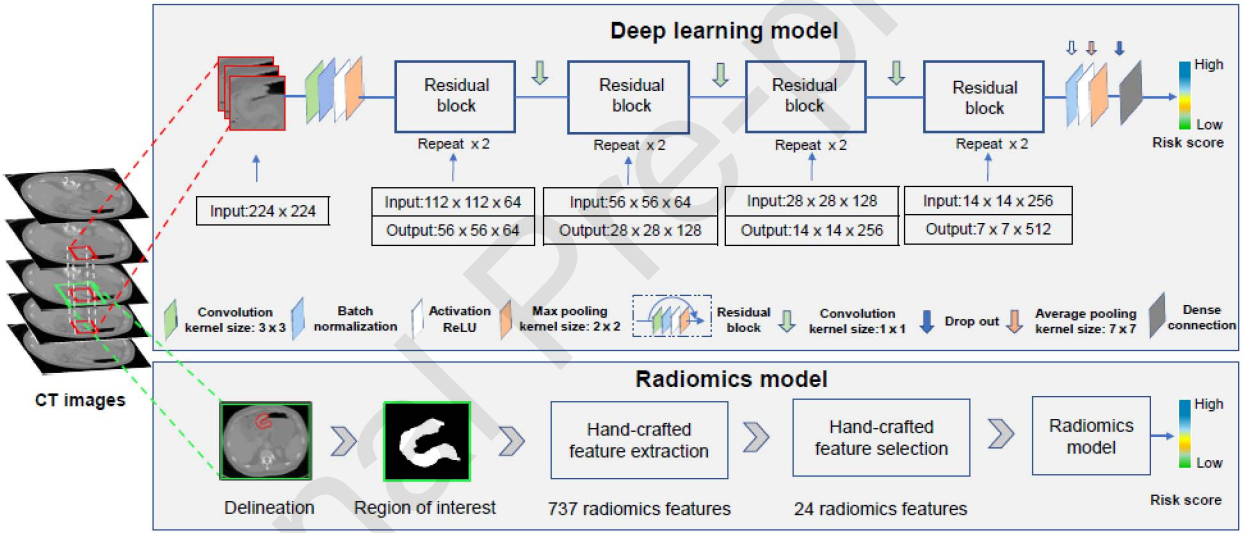
C

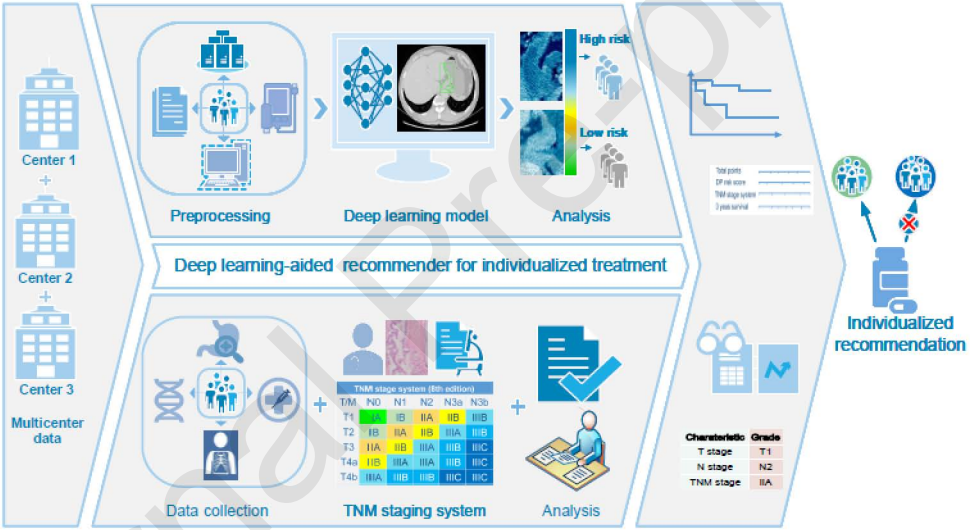
Model	C-index (95% CI)		Hazard ratio (95% CI)†	
	Training	External validation	Training	External validation
DL	0.82 (0.80-0.84)	0.78 (0.72-0.83)	9.79 (7.15-13.41)	11.76 (4.23-32.71)
Clinical‡	0.73(0.69-0.76)	0.71 (0.64-0.78)	3.84(2.91-5.07)	3.57(2.02-6.32)
Radiomics	0.66(0.63-0.69)	0.72(0.65-0.79)	2.48(1.91-3.22)	5.86(1.83-18.82)

D




A
Training
DL model
Strata Low High
Survival probability
p < 0.0001
Number at risk: n (%)
Low 259 (100) 239 (92) 166 (64) 21 (8)
High 259 (100) 88 (34) 8 (3) 0 (0)
Number of censoring
n.censor
Time
B
Training
Clinical model
Strata Low High
Survival probability
p < 0.0001
Number at risk: n (%)
Low 250 (100) 202 (81) 128 (51) 16 (6)
High 268 (100) 125 (47) 46 (17) 5 (2)
Number of censoring
n.censor
Time
C
Training
Radiomics model
Strata Low High
Survival probability
p < 0.0001
Number at risk: n (%)
Low 259 (100) 191 (74) 118 (46) 13 (5)
High 259 (100) 136 (53) 56 (22) 8 (3)
Number of censoring
n.censor
Time
D
External validation
DL model
Strata Low High
Survival probability
p < 0.0001
Number at risk: n (%)
Low 46 (100) 44 (96) 38 (83) 15 (33) 3 (7)
High 76 (100) 54 (71) 34 (45) 11 (14) 3 (4)
Number of censoring
n.censor
Time
E
External validation
Clinical model
Strata Low High
Survival probability
p < 0.0001
Number at risk: n (%)
Low 57 (100) 62 (93) 50 (75) 18 (27) 3 (4)
High 55 (100) 36 (65) 22 (40) 8 (15) 3 (5)
Number of censoring
n.censor
Time
F
External validation
Radiomics model
Strata Low High
Survival probability
p = 0.00075
Number at risk: n (%)
Low 25 (100) 24 (96) 23 (92) 8 (32) 1 (4)
High 97 (100) 74 (76) 49 (51) 18 (19) 5 (5)
Number of censoring
n.censor
Time


Deep learning model
Residual block
Repeat x 2
Residual block
Repeat x 2
Residual block
Repeat x 2
Residual block
Repeat x 2
High
Low
Risk score
Input:224 x 224
Input:112 x 112 x 64
Output:56 x 56 x 64
Input:56 x 56 x 64
Output:28 x 28 x 128
Input:28 x 28 x 128
Output:14 x 14 x 256
Input:14 x 14 x 256
Output:7 x 7 x 512
Convolution kernel size: 3 x 3
Batch normalization
Activation ReLU
Max pooling kernel size: 2 x 2
Residual block
Convolution kernel size:1 x 1
Drop out
Average pooling kernel size: 7 x 7
Dense connection
CT images
Radiomics model
Delineation
Region of interest
Hand-crafted feature extraction
737 radiomics features
Hand-crafted feature selection
24 radiomics features
Radiomics model
High
Low
Risk score


Center 1
Center 2
Center 3
Multicenter data
Preprocessing
Deep learning model
High risk
Low risk
Analysis
Deep learning-aided recommender for individualized treatment
Data collection
TNM staging system
Analysis
TNM stage system (8th edition)
T/M N0 N1 N2 N3a N3b
T1 IA IB IIA IIB IIIB
T2 IB IIA IIB IIIA IIIB
T3 IIA IIB IIIA IIIB IIIC
T4a IIB IIIA IIIA IIIB IIIC
T4b IIIA IIIB IIIB IIIC IIIC
Charasteristic Grade
T stage T1
N stage N2
TNM stage IIA
Individualized recommendation