Original Article

Predicting the recurrence risk of pancreatic neuroendocrine neoplasms after radical resection using deep learning radiomics with preoperative computed tomography images

Chenyu Song[1#], Mingyu Wang[2#], Yanji Luo[1#], Jie Chen[3], Zhenpeng Peng[1], Yangdi Wang[1], Hongyuan Zhang[2], Zi-Ping Li[1], Jingxian Shen[4], Bingsheng Huang[2], Shi-Ting Feng[1]

[1]Department of Radiology, The First Affiliated Hospital, Sun Yat-Sen University, Guangzhou, China; [2]Medical AI Lab, School of Biomedical Engineering, Health Science Center, Shenzhen University, Shenzhen, China; [3]Department of Gastroenterology, The First Affiliated Hospital, Sun Yat-sen University, Guangzhou, China; [4]Department of Radiology, Sun Yat-sen University Cancer Center, Guangzhou, China

Contributions: (I) Conception and design: ST Feng, B Huang; (II) Administrative support: ST Feng; (III) Provision of study materials or patients: All authors; (IV) Collection and assembly of data: All authors; (V) Data analysis and interpretation: C Song, M Wang, Y Luo; (VI) Manuscript writing: All authors; (VII) Final approval of manuscript: All authors.

[#]These authors contributed equally to this work.

Correspondence to: Jingxian Shen. Department of Radiology, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, 651 Dongfeng East Road, Guangzhou 510060, China. Email: shenjx@sysucc.org.cn; Bingsheng Huang. Medical AI Lab, School of Biomedical Engineering, Health Science Center, Shenzhen University. Block A2, Xili Campus of Shenzhen University, 1066 Xueyuan Avenue, Shenzhen 518073, China. Email: huangb@szu.edu.cn; Shi-Ting Feng. Department of Radiology, The First Affiliated Hospital, Sun Yat-Sen University, 58th, The Second Zhongshan Road, Guangzhou, Guangdong 510080, China. Email: fengsht@mail.sysu.edu.cn.

Background: To establish and validate a prediction model for pancreatic neuroendocrine neoplasms (pNENs) recurrence after radical surgery with preoperative computed tomography (CT) images.

Methods: We retrospectively collected data from 74 patients with pathologically confirmed pNENs (internal group: 56 patients, Hospital I; external validation group: 18 patients, Hospital II). Using the internal group, models were trained with CT findings evaluated by radiologists, radiomics, and deep learning radiomics (DLR) to predict 5-year pNEN recurrence. Radiomics and DLR models were established for arterial (A), venous (V), and arterial and venous (A&V) contrast phases. The model with the optimal performance was further combined with clinical information, and all patients were divided into high- and low-risk groups to analyze survival with the Kaplan-Meier method.

Results: In the internal group, the areas under the curves (AUCs) of DLR-A, DLR-V, and DLR-A&V models were 0.80, 0.58, and 0.72, respectively. The corresponding radiomics AUCs were 0.74, 0.68, and 0.70. The AUC of the CT findings model was 0.53. The DLR-A model represented the optimum; added clinical information improved the AUC from 0.80 to 0.83. In the validation group, the AUCs of DLR-A, DLR-V, and DLR-A&V models were 0.77, 0.48, and 0.64, respectively, and those of radiomics-A, radiomics-V, and radiomics-A&V models were 0.56, 0.52, and 0.56, respectively. The AUC of the CT findings model was 0.52. In the validation group, the comparison between the DLR-A and the random models showed a trend of significant difference (P=0.058). Recurrence-free survival differed significantly between high- and low-risk groups (P=0.003).

Conclusions: Using DLR, we successfully established a preoperative recurrence prediction model for pNEN patients after radical surgery. This allows a risk evaluation of pNEN recurrence, optimizing clinical decision-making.

Keywords: Pancreatic neuroendocrine neoplasms (pNENs); deep learning radiomics (DLR); survival analysis

Submitted Jan 04, 2021. Accepted for publication Mar 21, 2021.
doi: 10.21037/atm-21-25
View this article at: http://dx.doi.org/10.21037/atm-21-25

Introduction

Pancreatic neuroendocrine neoplasms (pNENs) are tumors with complex biological behaviors (1,2). R0 surgical resection is the first-line therapy for non-metastatic neuroendocrine neoplasms, but its postoperative recurrence is variable and difficult to predict, with 5-year recurrence rates ranging from 5% to 80% (3-6). If the probability of a pNEN recurrence (including local recurrence and distant metastasis) could be accurately predicted before surgery, the preoperative surgical plan could be optimized, and the management of the postoperative follow-up and intervention could be arranged in advance. This strategy can minimize the recurrence probability and reduce the adverse impact of postoperative tumor recurrence, thus improving the prognosis of patients (7). Specifically, for patients with low recurrence risk, the frequency of surveillance can be reduced and a relatively longer monitoring interval can be set (8). For patients at high risk of recurrence, surgical margins should be expanded, and lymph node dissection should be more thorough in their preoperative surgical plan. Likewise, emphasis should be placed on postoperative follow-up and combined treatment in these patients.

Some studies reported methods for recurrence prediction in pNENs (1,9). Pathological parameters including the Ki-67 index of postoperative specimens or preoperative biopsy were used to predict the prognosis of pNEN patients (10,11). However, these studies were either based on indicators obtained after surgery or from fractional tissue, and the predictive performance was unsatisfactory with sensitivity (SEN) values of less than 40% including the Ki-67 index. Thus, these approaches cannot effectively guide preoperative management.

Computed tomography (CT) is commonly used for the diagnosis of pancreatic diseases with high diagnostic accuracy (ACC) in pNENs. Several studies (12-14) have shown that CT findings such as tumor size, tumor vascularity, and CT value can be used to predict the prognosis preoperatively. A CT ratio (the CT value of the tumor divided by that of non-tumorous pancreatic parenchyma) <0.85 and tumor size ≥3.0 cm were shown to be independent prognostic factors associated with the disease-free survival of patients with pNEN (14). However, these studies are all based on indicators evaluated by radiologists with inevitable subjectivity and measurement errors. In addition, such studies are generally limited, because relevant indicators are only for factor analysis but not used for the establishment and validation of more practical prediction models.

Radiomics has achieved great success in medical image analysis. Image features with strong identification power can be automatically analyzed with high throughput and extracted by computers for auxiliary diagnosis or therapy response prediction. In research, radiomics has been a commonly used method to predict the prognosis of patients, and medical image analysis technology based on deep learning brings more opportunities and challenges for prognosis prediction. Wang *et al.* (15) achieved ^{18}F-fluorodeoxyglucose (FDG) positron emission tomography (PET)/CT image-based prediction of lymph node metastasis in non-small cell lung cancer with deep learning. Using deep learning techniques, another group (16) established a 3-year recurrence prediction model for patients with ovarian cancer based on CT images. Chen *et al.* (17) used enhanced CT images and other predictors (tumor location, size, and other information) to establish the ResNet model that predicts the 3- and 5-year recurrence-free survival (RFS) rates for gastrointestinal stromal tumor patients with area under the curve (AUC) values of more than 0.90. A newly developed method called deep learning radiomics (DLR) (18,19) can extract quantitative and high-throughput features from medical images by pretrained artificial neural networks. This approach is different from the radiomics method that extracts explicitly designed features, and DLR has been proved a promising tool for computer-aided tumor prognosis prediction. It has been successfully applied to many clinical problems, such as predicting the stages of liver fibrosis (18) and predicting axillary lymph node metastasis in early-stage breast cancer (20).

To our knowledge, no study has assessed the recurrence prediction in pNEN patients based on radiomics or DLR techniques yet. This study aimed to establish and validate a recurrence prediction model for pNEN patients after radical surgery based on their preoperative CT images using CT findings evaluated by radiologists, radiomics, and DLR. We present the following article in accordance with the STROBE reporting checklist (available at http://dx.doi.org/10.21037/atm-21-25).

Methods

Study design

The study design is shown in *Figure 1*. We separately extracted features for training models based on the data from Hospital I (the First Affiliated Hospital of Sun Yat-Sen University, internal group) in three ways, namely, CT findings were evaluated by radiologists, radiomics, and DLR.

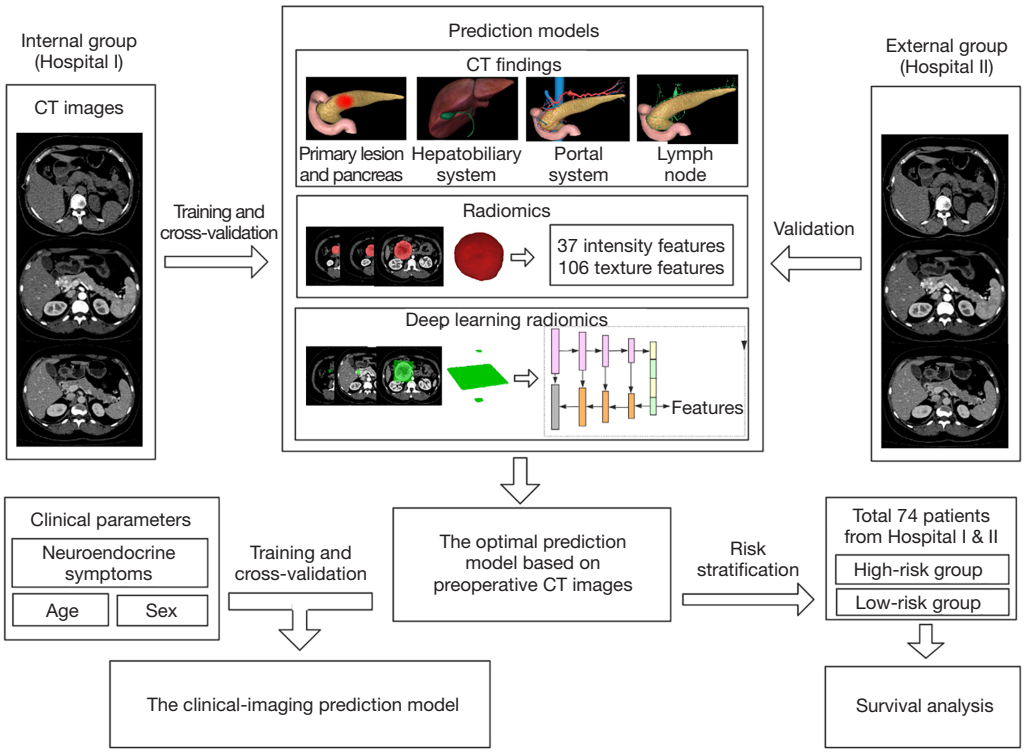

Figure 1 Flow chart of the study design. Computed tomography (CT) images were obtained in the unenhanced, arterial, and venous phases. Data from Hospital I were used to establish the prediction models (radiologist assessment, radiomics, and deep learning radiomics). Then, the external group from Hospital II was used to validate the prediction models. After the optimal prediction model had been selected, clinical indicators were added to observe changes in the predictive performance of this optimal model. In addition, an optimum model-based risk stratification model was established to explore its survival predictive potential.

Among them, radiomics and DLR were both used to extract features from images in the arterial, venous, and arterial & venous phases. In the second step, the models were validated. After cross-validation was completed using the internal group, an external validation was performed with CT images from Hospital II (Sun Yat-sen University Cancer Center, independent external group). Afterward, clinical indicators were added to the optimal model, and cross-validation was again performed on the data from Hospital I to observe the impact of clinical indicators on the predictive performance of the optimal model. In the last step, we constructed an optimum model-based risk stratification model to explore its survival predictive potential.

Acquisition of patient data

Patient selection and clinical data

This study was conducted in strict accordance with the principles of the Declaration of Helsinki (as revised in 2013). This retrospective study was approved by the Institutional Review Board of the First Affiliated Hospital of Sun Yat-sen University (No.: 2018-181), and written informed consent was waived by the Institutional Review Board. All patients had pNENs, that were pathologically confirmed after radical surgery from Hospital I and Hospital II, between 2010 and 2018 and did not receive any drugs or surgical treatment at the time of (or before) CT imaging. Patients with one of the following four conditions were excluded from the study population: (I) distant metastases had been detected in their first examination; (II) another concomitant malignancy was diagnosed; (III) a multiple endocrine neoplasia syndrome was confirmed; and (IV) not all CT images were available. The data filtering process is shown in *Figure 2* and the data inclusion and exclusion criteria of the two medical centers were consistent.

This study included three clinical parameters: age, sex, and

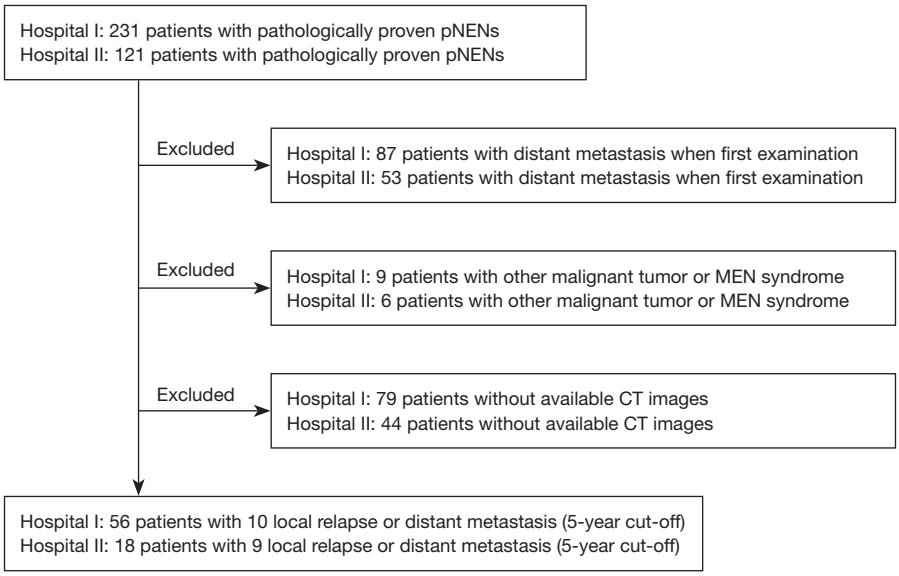

Figure 2 Data filtering procedure. pNEN, pancreatic neuroendocrine neoplasm; MEN, multiple endocrine neoplasia; CT, computed tomography.

neuroendocrine symptoms. Neuroendocrine symptoms were defined as relevant symptoms typically caused by excessive secretion of hormones in patients with corresponding elevated hormone levels detected in blood samples. Patients were followed up from the date of surgery to May 24, 2019. A medical imaging examination [ultrasound/CT/magnetic resonance imaging (MRI)] was performed at least once every 6 months in the first year, and once every six months or 1 year according to tumor pathological grade after 1 year (G1: once every year, G2/3 or neuroendocrine carcinoma: once every 6 months). PET-CTs with ^{68}Ga-labeled somatostatin analogues and ^{18}F-labeled FDG were used to examine suspected cases of recurrence. The date of recurrence (including local recurrence and distant metastasis) was defined as the time of recurrence detected by cross-sectional imaging (CT/MRI) during the follow-up. The neoplasm grew at the primary site or other organ confirmed by PET-CT or biopsy was defined as local recurrence or distant metastasis.

For pNEN patients with postoperative recurrence, the 5-year RFS was defined as the time from the date of surgery to the date of the first detection of a postoperative recurrence. For patients without postoperative recurrence, the RFS was defined as the time from the date of surgery to the date of the latest follow-up.

CT image acquisition

CT scans were performed in Hospital I using a 64-slice spiral CT scanner (Aquilion 64; Canon Medical Systems). The scanning parameters were as follows: 0.5-mm slice thickness, 0.5-mm slice interval, 200-mAs tube current, and 120-kVp tube voltage. An iodinated contrast agent (Ultravist 300; Bayer Schering, Berlin) was administered intravenously at a rate of 3 mL/s via a high-pressure syringe after pre-contrast imaging followed by a saline chaser bolus (40 mL) at the same rate. The arterial and venous phases were obtained at 35 and 65 s after contrast injection, respectively.

In Hospital II, the CT images were captured using a 128-slice spiral CT system (Discovery CT750 HD; GE System, Milwaukee, WI, USA). The scanning parameters were as follows: 2-mm slice thickness, 1-mm slice interval, automatic tube current modulation (maximum 450 mAs), and 100–140-kVp tube voltage. The contrast agent was administered as described for Hospital I. The arterial and venous phases were obtained after the aortic opacification reached 100 Hounsfield units (HU). The average scan started after contrast injection at 36 s (range, 30–42 s) for the arterial phase and 66 s (range, 58–70 s) for the venous phase.

CT image analysis

CT findings

Regarding the CT findings assessed by radiologists, the conditions of the primary lesion, pancreas, lymph node,

hepatobiliary system, and portal system were all evaluated independently by two radiologists with more than 10 years of experience in the imaging diagnosis of abdominal diseases that were blinded to the patients' pathological results. A detailed description of evaluated CT findings is shown in Table S1.

Radiomics

For radiomics, regions of interest (ROIs) were delineated by two radiologists that were responsible for the CT image evaluation, and they were also blinded to patients' pathological results during the whole process. The ground truth (GT) values of all patients were labeled on CT images in arterial and venous phases using the ITK-SNAP software, as shown in Figure S1.

First, we converted the raw data from the DICOM to the NIFTI format. According to the experience of the radiologists, the window level and window width were set for the arterial phase to 130 and 310 HU, respectively, and for the venous phase to 120 and 320 HU, respectively. Finally, the voxel size of all images was resampled to 1 mm × 1 mm × 1 mm using the 3D cubic interpolation algorithm.

Based on our own developed toolkit, we extracted 143 features describing intensity [37] and texture [106], such as gray-level co-occurrence matrix, spatial gray-level dependence matrix, neighborhood gray-tone difference matrix, and neighborhood gray-level difference statistics.

DLR

The prediction process based on DLR required only rough annotations by the radiologists. The ROI annotation was also completed by the two radiologists who were responsible for the CT image evaluation, and the pathological results of the patients were not disclosed to the radiologists. Each radiologist delineated the top layer, the largest layer, and the bottom layer of each tumor in the cross section. No strict criteria were used for delineation. The radiologists had only to draw the quadrilateral area containing the tumor area, as shown in Figure S1.

We additionally collected 58 CT images without recurrence tags but with the GT of the segmentation in the arterial phase, and the scanning parameters of these data were consistent with those of the Hospital I data. These data were used specifically for the training and validation of the segmentation network. To ensure that the network learned the characteristics with higher identification power, we randomly sampled 22 cases from the Hospital I data set and mixed them with the additionally collected 58 cases, to randomly divide them at a 1:9 ratio into a verification and a training set. Then, we trained a two-dimensional U-net to extract DLR features. Further details are presented in Figure S2 and the Supplementary Materials and Methods ("Training U-net for DLR" section).

Data preprocessing was performed as described in the "Traditional radiomics" section. The data for the training network only comprised the arterial phase GT, and both arterial and venous phase images used the pretrained arterial model when extracting features. We included all slices into the pretrained U-net to retrieve slice-wise features, and then, we used a clustering-based method to aggregate the slice-wise features into patient-wise features. Details of the feature extraction procedure are shown in the Supplementary Materials and Methods ("DLR features extraction" section).

Training and validation of the prediction models

Training and cross-validation of the models

We built the recurrence prediction models using a support vector machine algorithm (based on Scikit-learn machine learning library). We used 10-fold cross-validation on the internal group to evaluate the performance of the prediction model. In each fold, the two-sample *t*-test or Mann-Whitney U test was performed, and the features which showed significant differences between recurrence *vs.* recurrence-free groups were selected from the training set before applying the selection results to the test set. The model parameters for each fold were determined using the grid searching method on the training set. The main evaluation indicators of the final model were ACC, SEN, specificity (SPC), and AUC. The receiver operating characteristic (ROC) curves of all models were compared with the random case (AUC =0.5), and the AUC values between models were also compared using the DeLong test performed by MedCal software (version 12.5.0.0 by MedCal software bvba).

External independent validation

Using the ROC curves of the external independent dataset, we evaluated the robustness of the method which provided the optimal performance on internal dataset. We used a model integration approach to predict recurrence risk in external independent validation. The details of the model integration are presented in the Supplementary Materials and Methods ("Model integration" section).

Statistical analysis

Clinical information and CT findings were analyzed using

Table 1 Accuracy, sensitivity, specificity, and AUC values of the radiomics models for recurrence prediction (56 patients from Hospital I and 18 patients from Hospital II)

Models	Hospital I (internal data set)					Hospital II (independent data set)				
	ACC	SEN	SPC	AUC	P	ACC	SEN	SPC	AUC	P
Radiomics-A	0.75	0.70	0.76	0.74	0.020	0.44	0.11	0.78	0.56	0.691
Radiomics-V	0.71	0.80	0.70	0.68	0.083	0.44	0.33	0.56	0.52	0.965
Radiomics-A&V	0.71	0.80	0.70	0.70	0.044	0.56	0.22	0.89	0.56	0.691

The threshold of the predictive probability used to calculate ACC, SEN, and SPC was the highest Youden index of the cross-validation ROC curves for the internal data set. A P value indicates the significance level of the comparison between an AUC with that of a random case (AUC =0.5). AUC, area under the curve; ACC, accuracy; SEN, sensitivity; SPC, specificity; A, arterial; V, venous; A&V, arterial & venous.

univariate analysis. Continuous variables conforming to normal distribution were described by the mean and standard deviation, and the independent two-sample *t*-test was performed. If a normal distribution was not confirmed, the median and interquartile ranges were used, and the Mann-Whitney U test was performed as a non-parametric test. For categorical variables, the χ^2 test or exact probability method was used in this study. $P<0.05$ was defined as statistically significant. In this study, based on the optimal prediction model, patients from the two hospitals were divided into high- and low-risk groups for Kaplan-Meier analyses. Patients were stratified into high- and low-risk groups using the threshold of predicted recurrence probability defined as the highest Youden index (21) of the cross-validation ROC curves. All statistical analysis were performed by SPSS software (version 25.0 for Macintosh, IBM, Chicago, IL, USA).

Results

Finally, a total of 74 pNEN patients were included in this study. Fifty-six patients (recurrence of 10 patients within 5 years) of Hospital I are used for training and internal validation, and 18 patients (recurrence of 9 patients within 5 years) of Hospital II are used for external independent validation. The clinical information of patients of Hospital I is shown in Table S2 with CT findings. As for patients of Hospital II, neither the sex (6 females with recurrence of 3 patients, 12 males with recurrence of 6 patients) nor the mean age (53.56±10.36 in recurrence group, 48.78±15.67 in recurrence-free group) was significantly different between the recurrence and recurrence-free groups.

We annotated (by C Song and Y Luo with 4 and 8 years of working experience, respectively) on 5 random cases from the data. The mean time of the two radiologists to locate were 11.30 and 9.98 s, and the medians were 11.04 and 9.79 s. The two radiologists spent an average of 647.19 and 796.01 s in the fine-delineation process, with a median of 305.51 and 382.59 s, respectively.

Clinical information and CT findings

The results of the univariate analysis for Hospital I are shown in Table S2. Among the examined factors, the CT ratios of the primary lesion in the unenhanced phase and the venous phase were significantly different between the recurrence and recurrence-free groups. Neuroendocrine symptoms, the shape and size of the primary lesion, the shape of pancreatic duct, lymph node morphology, and lymph node enhancement pattern were all significantly associated with recurrence. There were more patients with tumor recurrence in the groups with asymptomatic tumors, cystic-solid tumors, tumors with a maximum diameter greater than 20 mm, the dilation or cutoff of pancreatic duct, normal lymph node size, and homogeneous lymph node enhancement. In the univariate analysis for Hospital II, among the examined CT findings, only the CT ratio and the relatively enhanced rate of the primary lesion in the arterial phase and the venous phase were significantly different between the recurrence and recurrence-free groups. The AUCs were 0.53 and 0.52 respectively in the internal and validation groups.

Radiomics

Table 1 shows the results of the 10-fold cross-validation of the radiomics model based on features in different phases extracted from the data of Hospital I. Using the data of

Table 2 Accuracy, sensitivity, specificity, and AUC values of the DLR models for recurrence prediction (56 patients from Hospital I and 18 patients from Hospital II)

Models	Hospital I (internal data set)					Hospital II (independent data set)				
	ACC	SEN	SPC	AUC	P	ACC	SEN	SPC	AUC	P
DLR-A	0.71	0.90	0.67	0.80	0.003	0.61	0.55	0.66	0.77	0.058
DLR-V	0.73	0.60	0.76	0.58	0.429	0.44	0.22	0.67	0.48	0.895
DLR-A&V	0.71	0.80	0.70	0.72	0.034	0.61	0.44	0.78	0.64	0.310

The threshold of the predictive probability used to calculate ACC, SEN, and SPC was the highest Youden index of the cross-validation ROC curves for the internal data set. A P value indicates the significance level of the comparison between an AUC with that of a random case (AUC =0.5). AUC, area under the curve; DLR, deep learning radiomics; ACC, accuracy; SEN, sensitivity; SPC, specificity; A, arterial; V, venous; A&V, arterial & venous.

Table 3 Performance comparison between the optimal radiomics model (radiomics-A), the optimal DLR model (DLR-A), and the model based on CT findings (56 patients from Hospital I)

Model	ACC	SEN	SPC	AUC	P
Radiomics-A	0.75	0.70	0.76	0.74	0.020
DLR-A	0.71	0.90	0.67	0.80	0.003
CT findings	0.63	0.50	0.65	0.53	0.748

A P value indicates the significance level of the comparison between an AUC with that of a random case (AUC =0.5). DLR, deep learning radiomics; A, arterial; CT, computed tomography; ACC, accuracy; SEN, sensitivity; SPC, specificity; AUC, area under the curve.

Hospital II, the results of the established prediction models were verified independently. According to the cross-validation and independent validation results, the model performed best in the arterial phase with an AUC of 0.74 for cross-validation and 0.56 for independent validation. The results of the DeLong test comparing the different phases are shown in Table S3. There were no significant differences in AUCs for different phases. The ROC curves of the radiomics model in different contrast phases are shown in Figure S4.

DLR

The 10-fold cross-validation results based on DLR features are shown in *Table 2*. The Hospital II data were used to validate this model independently. The model reached the highest AUCs in the arterial phase both for cross-validation (0.80) and independent validation (0.77). The ROCs are compared in Table S3. For the different phases, no significant differences were detected in the cross-validation results. Figure S4 shows the ROC curves of all models trained with DLR.

Optimal prediction model with and without added clinical features, the comparison of the optimal radiomics model, the optimal DLR model, and the model based on CT findings regarding the prediction of postoperative tumor recurrence is shown in *Table 3*. The highest cross-validated AUC value was observed in the DLR model of the arterial phase (DLR-A; AUC =0.80). The cross-validation results with added clinical information (not included in the feature extraction) are shown in *Table 4*. After including the three clinical parameters, all model indicators, except for SEN decreasing by 0.10, were improved to some extent with ACC, SPC, and AUC reaching 0.80, 0.80, and 0.83, respectively. However, the ROC results of the models before and after the addition of the clinical information were not significantly different. As shown in Table S3, all image-based models showed no statistically significant differences between each other. *Figure 3A* displays the ROC curves of the optimal radiomics model (radiomics-A), the optimal DLR model (DLR-A), and the model based on CT findings. The ROC curve of the DLR-A model with added clinical information is presented in *Figure 3B*.

Survival analysis

Using the predicted value of the DLR-A model as the risk factor and the highest Youden index in the internal group

Table 4 Accuracy, sensitivity, specificity, and AUC values of the DLR-A recurrence prediction model with added clinical information (56 patients from Hospital I)

Model	ACC	SEN	SPC	AUC	P[a]	P[b]		
						DLR-A + s	DLR-A + sa	DLR-A + sag
DLR-A	0.71	0.90	0.67	0.80	0.003	0.413	0.822	0.680
DLR-A + s	0.71	0.80	0.70	0.75	0.015	–	0.459	0.108
DLR-A + sa	0.76	0.90	0.73	0.79	0.004	–	–	0.483
DLR-A + sag	0.80	0.80	0.80	0.83	0.001	–	–	–

[a], a P value indicates the significance level of the comparison between an AUC with that of a random case (AUC =0.5). [b], a P value indicates the significance level of comparison between every two AUCs. AUC, area under the curve; DLR, deep learning radiomics; A, arterial; ACC, accuracy; SEN, sensitivity; SPC, specificity; + s, symptom added; + sa, symptom and age added; + sag, symptom, age, and gender added.

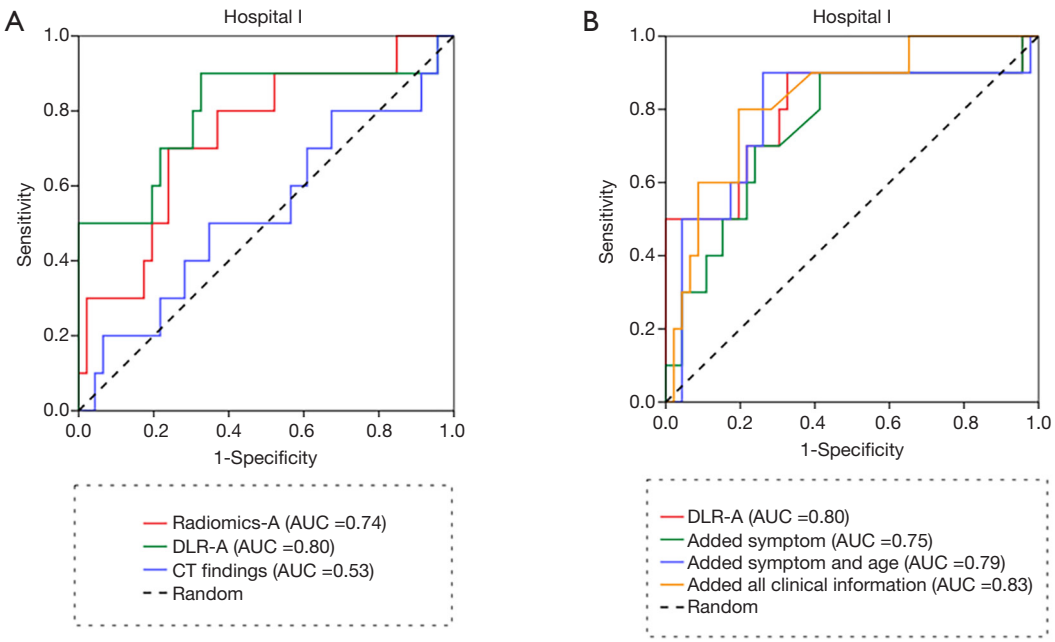

Figure 3 The receiver operating characteristic (ROC) of deep learning radiomics (DLR), radiomics and CT findings models in Hospital I. (A) The receiver operating characteristic (ROC) curves of the optimal radiomics (R) model (R-A), the optimal deep learning radiomics model (DLR-A), and the model based on CT findings. (B) The ROC curves of the DLR-A model with added clinical information. AUC, area under the curve.

as its stratification threshold (0.165499), the combined patients from both hospitals were divided into a high-risk and a low-risk group. The mean and median survival times were in the high-risk group 36.28 months [95% confidence interval (CI), 26.37 to 46.20 months] and 38.53 months (95% CI, 10.63 to 66.44 months), respectively. The mean survival time in the low-risk group was 53.11 months (95% CI, 46.91 to 59.32 months). The survival analysis using the Kaplan-Meier method is shown in *Figure 4*, in which the P value of the log-rank test is 0.003.

Discussion

In this study, we successfully established recurrence prediction models for pNEN patients based on three methods: radiologist assessment, radiomics, and DLR.

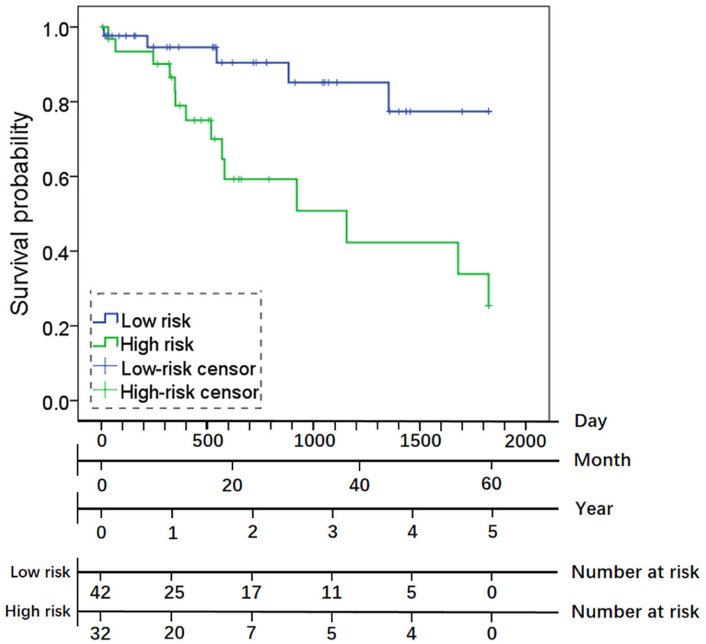

Figure 4 Survival analysis using the high- and low-risk groups according to the DLR-A model. The Kaplan-Meier analysis shows a statistically significant difference (P=0.003; log-rank test) between these groups regarding recurrence-free survival. DLR, deep learning radiomics; A, arterial.

We also analyzed the influence of CT imaging phase and clinical information on the performance of the prediction models. Compared with previous studies (1,9,12,22,23) on postoperative recurrence in pNEN patients, we found that most of these studies performed univariate analyses based on biochemical indicators or CT findings without an established and validated prediction model. Some indicators like the Ki-67 index or the pathological grade can only be obtained after surgery limiting their practical application. Our study applied radiomics successfully to postoperative recurrence prediction in patients with pNEN based on preoperative parameters. The DLR-A model performed optimally on both internal and external data sets but without significantly difference between the models. The results of the CT findings in our study were consistent with those in previous publications (13,14,24-26). Smaller and round lesions often indicated less aggressive behavior or early discovery, which are both associated with a better prognosis. The CT ratio represents the difference in CT values between the primary lesion and the pancreas parenchyma. In the unenhanced phase, more patients in the recurrence-free group showed lower attenuation of the primary lesion relative to the pancreas parenchyma. An explanation might be that these lesions contained fewer solid components or that tumor cells proliferated more slowly. pNENs are highly vascularized tumors. Thus, they were significantly enhanced in the arterial phase in both the recurrence and the recurrence-free groups. However, in the venous phase, more patients in the recurrence group presented a lower attenuation of the primary lesion relative to the pancreas parenchyma. Possibly, the lesions of the recurrence group contained more blood vessel connections leading to faster blood flow, and consequently, a more obvious CT value decrease in the venous phase would be observed in the recurrence group. Regarding the lymph node morphology, the confluent multinodular lymph node group with its 100% recurrence rate comprised only one patient. In the enlarged lymph node group and the normal lymph node group, 60% [3/5] and 15% [6/41] of the patients presented with postoperative recurrence, respectively. Considering the aspect of lymph node enhancement patterns, the two patients of the heterogeneous enhancement group both presented with postoperative recurrence. Lymph node enlargement, fusion, and heterogeneous enhancement are important indicators of lymph node metastasis in various tumor types, and the same is true for pNENs (27). Given

that the liver is the organ most susceptible to metastasis and that the venous reflux from the pancreas is drained through the portal system of the liver, the CT findings of the liver and its portal system were also included in our study. However, these group differences were not statistically significant. A larger pNEN sample size may be needed for further explorations.

The model based on CT findings required the manual image evaluation by the radiologists, whereas the radiomics model involved the radiologists precisely delineating the ROIs. Although in this study as in most previous studies experienced radiologists were employed to avoid the variability and subjectivity of manually delineated ROIs, the level of experience to evaluate CT images differs in practice. Our simple semi-automatic method used in the DLR prediction model greatly reduced subjectivity and task complexity, while achieving high SEN by roughly locating the tumor. Although the ACC and SPC values of the external independent validation were low due to some deviation in the distribution of features, the AUC value of these data reached 0.77 indicating that the model still had a robust ability for risk stratification. We used a segmented network-based DLR method using image properties (mask) to supervise the network training and to automatically obtain more force-expressing features with less data volume. Therefore, over- or underfitting problems due to the use of unbalanced recurrent tags to supervise the training were avoided. Another study (23) conducted by our research team used only grading labels for supervision, and the results demonstrated that the deep learning method was not superior to the radiomics approach. Moreover, the findings suggested that the use of semantic labels such as grading or recurrence labels to supervise networks might limit the performance with small training sets.

This study also compared the performance of the models based on different contrast enhancement phases, and we found that the models in the arterial phase performed superior to the models in the venous or arterial & venous phases for both radiomics and DLR models. This result is consistent with the previous findings of our team (23). The reasons for this are as follows: (1) Most pNEN lesions are highly vascularized. Thus, the difference between the primary tumor and the surrounding normal pancreas parenchyma is more obvious in the arterial phase, and the tumor outline is more clearly displayed. By contrast, the demarcation between tumor and surrounding parenchyma is relatively poor in the venous phase. The segmentation network would, therefore, better acquire the ability to distinguish the tumor from the surrounding tissue in the arterial phase. In other words, the segmentation network excluded any interference from the surrounding tissue in the arterial phase and could pay more attention to the characteristics of the tumor itself (2). Compared to DLR, the feature extraction was in the radiomics model limited by the radiologist-defined tumor boundaries. Because pNENs are better distinguishable in the arterial phase, it was easier to observe characteristics such as texture in the arterial than in the venous phase (3). In the DLR models, the characteristics with poor performance may be due to the obscured tumor contour in the venous phase and the inability of the network to effectively identify the tumor area. That a network trained on arterial data was unsuitable for venous data may be another reason, and the inherent phase differences led to the transfer failure (4). Finally, the performances of the combined arterial & venous phase models were for both DLR and radiomics methods not as good as the arterial phase models, which may have been caused by feature redundancy. Feature redundancy means that features have a high degree of collinearity. The same situation occurred in our previous study (23). Theoretically, high collinearity can lead to poor model prediction performance (28). We performed a collinearity analysis on the DLR features of arterial phase and venous phase, and the results showed that most of the two features have a high degree of collinearity (Figure S5). Therefore, the redundant information brought by the highly collinearity feature is the reason why the DLR-arterial & venous (DLR-A&V) model is inferior to the DRL-A model. In the current study we added clinical information to the optimal DLR-A model and found that the performance was improved without reaching statistical significance. This indicates the importance of clinical information and its positive effects on the modeling process.

In this study, the optimal model (DLR-A) was selected to stratify the risk of postoperative recurrence in pNEN patients from two hospitals. According to the results of the Kaplan-Meier analysis, in the DLR-A model that determined the recurrence probability with a 5-year RFS cutoff, the survival rates differed significantly between high- and low-risk groups. Moreover, the survival analysis included not only the final status of a patient but also information about the time to reach this status. Compared with other model evaluation indicators like AUC, ACC, etc., the results of the survival analysis (such as survival time, mean survival time) reflected the ability of the prediction model to stratify the survival status of patients and their

theoretical survival status with more practical significance.

In our study, none of the models performed was better than random chance in external data set. There may be heterogeneity in the imaging data due to the different parameters in the scanning process at different centers, which can reduce the generalization ability of the prediction models. As indicated in some recent studies (29-31), domain adaptive technology based on deep learning may be applied to reduce the difference in data distribution to improve the generalization ability of the method in further studies. Another factors, such as surgeons of different experience in different hospitals, and postoperative monitoring frequency, etc., can be influential, and require prospective studies to verify.

There are some limitations to our study. First, although the DLR-A model was the optimal model in our study, it was still only a semi-automatic method that requires a radiologist to provide information regarding the tumor location. Fully automatic localization or segmentation for feature extraction is warranted, not only to avoid the subjectivity of a radiologist but also to improve the prediction performance. Second, similar to our study, published studies in patients with pNENs are mostly limited by small data sets (23,32,33). This might be the reason that no statistically significant differences within each model were detected. However, the difference between the DLR-A and the random model was statistically significant in the internal group and in the external group, the difference between the DLR-A and the random model was nearly significant. The survival analysis also demonstrated the potential of the optimal model for prognosis prediction. In this study, the independent dataset was small. It needs a larger external dataset to further prove the robustness of the model in the future. Third, in four patients of the external group, the records regarding their neuroendocrine symptoms were not available. Therefore, we failed to validate the model with added clinical features using the data from Hospital II. Fourth, in our presented study, we did not predict two outcomes (local recurrence and distal metastasis) separately due to the limitation of the sample size. Finally, the performance of our optimal model remains to be improved with emerging artificial intelligence technologies. We believe that these technologies can overcome the problems of sample size and annotation to further improve the ACC of the prediction model.

Conclusions

In summary, this study successfully established a preoperative prediction model of pNEN recurrence with good generalization in an external data set. It provides the basis to evaluate the risk of postoperative recurrence in pNEN patients with high SEN, thus aiding decision-making processes in clinical practice. But how individual follow-up surveillance and treatment plans in patients with different postoperative risks should be performed, needs to be further explored based on the results of the current study.

Acknowledgments

Funding: This work was funded by National Natural Science Foundation of China (81571750, 81771908, 81971684); Shenzhen-Hong Kong Institute of Brain Science-Shenzhen Fundamental Research Institutions (2019SHIBS0003); Tencent "Rhinoceros Birds"-Scientific Research Foundation for Young Teachers of Shenzhen University; 2020 SKY Imaging Research Fund of the Chinese International Medical Foundation (Z-2014-07-2003-07); Shenzhen Science and Technology Project (JCYJ20200109114014533); SZU Top Ranking Project, Shenzhen University (860/000002100108); Guangdong Basic and Applied Basic Research Foundation, GRANT (2020A1515010571); Seed Funding from Guangzhou Science and Technology Planning Project (201903010073); Guangdong College Students' Science and Technology Innovation Cultivation Project (pdjh2020a0497).

Footnote

Reporting Checklist: The authors have completed the STROBE reporting checklist. Available at http://dx.doi.org/10.21037/atm-21-25

Data Sharing Statement: Available at http://dx.doi.org/10.21037/atm-21-25

Peer Review File: Available at http://dx.doi.org/10.21037/atm-21-25

Conflicts of Interest: All authors have completed the ICMJE uniform disclosure form (available at http://dx.doi.org/10.21037/atm-21-25). The authors have no conflicts of interest to declare.

Ethical Statement: The authors are accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are

appropriately investigated and resolved. This study was conducted in strict accordance with the principles of the Declaration of Helsinki (as revised in 2013). This retrospective study was approved by the Institutional Review Board of the First Affiliated Hospital of Sun Yat-sen University (No: 2018-181), and written informed consent was waived by the Institutional Review Board.

References

1. Marchegiani G, Landoni L, Andrianello S, et al. Patterns of recurrence after resection for pancreatic neuroendocrine tumors: Who, when, and where? Neuroendocrinology 2019;108:161-71.
2 Tai WM, Tan SH, Tan DMY, et al. Clinicopathologic characteristics and survival of patients with gastroenteropancreatic neuroendocrine neoplasm in a multi-ethnic Asian institution. Neuroendocrinology 2019;108:265-77.
3 Masui T, Sato A, Nakano K, et al. Comparison of recurrence between pancreatic and duodenal neuroendocrine neoplasms after curative resection: A single-institution analysis. Ann Surg Oncol 2018;25:528-34.
4 Turaga KK, Kvols LK. Recent progress in the understanding, diagnosis, and treatment of gastroenteropancreatic neuroendocrine tumors. CA Cancer J Clin 2011;61:113-32.
5 Dieckhoff P, Runkel H, Daniel H, et al. Well-differentiated neuroendocrine neoplasia: Relapse-free survival and predictors of recurrence after curative intended resections. Digestion 2014;90:89-97.
6 Ter-Minassian M, Chan JA, Hooshmand SM, et al. Clinical presentation, recurrence, and survival in patients with neuroendocrine tumors: results from a prospective institutional database. Endocr Relat Cancer 2013;20:187-96.
7 Liu DJ, Fu XL, Liu W, et al. Clinicopathological, treatment, and prognosis study of 43 gastric neuroendocrine carcinomas. World J Gastroenterol 2017;23:516-24.
8 Pulvirenti A, Javed AA, Landoni L, et al. Multi-institutional Development and External Validation of a Nomogram to Predict Recurrence After Curative Resection of Pancreatic Neuroendocrine Tumors. Ann Surg 2019. [Epub ahead of print]. doi:10.1097/SLA.0000000000003579.
9 Feng T, Lv W, Yuan M, et al. Surgical resection of the primary tumor leads to prolonged survival in metastatic pancreatic neuroendocrine carcinoma. World J Surg Oncol 2019;17:54.
10 Genç CG, Falconi M, Partelli S, et al. Recurrence of pancreatic neuroendocrine tumors and survival predicted by Ki67. Ann Surg Oncol 2018;25:2467-74.
11 Liang W, Yang P, Huang R, et al. A combined nomogram model to preoperatively predict histologic grade in pancreatic neuroendocrine tumors. Clin Cancer Res 2019;25:584-94.
12 Shen C, Dasari A, Chu Y, et al. Clinical, pathological, and demographic factors associated with development of recurrences after surgical resection in elderly patients with neuroendocrine tumors. Ann Oncol 2019;30:1847.
13 Yamada S, Fujii T, Suzuki K, et al. Preoperative identification of a prognostic factor for pancreatic neuroendocrine tumors using multiphase contrast-enhanced computed tomography. Pancreas 2016;45:198-203.
14 Yamamoto Y, Okamura Y, Uemura S, et al. Vascularity and tumor size are significant predictors for recurrence after resection of a pancreatic neuroendocrine tumor. Ann Surg Oncol 2017;24:2363-70.
15 Wang H, Zhou Z, Li Y, et al. Comparison of machine learning methods for classifying mediastinal lymph node metastasis of non-small cell lung cancer from 18F-FDG PET/CT images. EJNMMI Res 2017;7:11.
16 Wang S, Liu Z, Rong Y, et al. Deep learning provides a new computed tomography-based prognostic biomarker for recurrence prediction in high-grade serous ovarian cancer. Radiother Oncol 2019;132:171-7.
17 Chen T, Liu S, Li Y, et al. Developed and validated a prognostic nomogram for recurrence-free survival after complete surgical resection of local primary gastrointestinal stromal tumors based on deep learning. EBioMedicine 2019;39:272-9.
18 Wang K, Lu X, Zhou H, et al. Deep learning Radiomics of shear wave elastography significantly improved diagnostic performance for assessing liver fibrosis in chronic hepatitis

B: a prospective multicentre study. Gut 2019;68:729-41.
19 Gillies RJ, Kinahan PE, Hricak H. Radiomics: Images are more than pictures, they are data. Radiology 2016;278:563-77.
20 Zheng X, Yao Z, Huang Y, et al. Deep learning radiomics can predict axillary lymph node status in early-stage breast cancer. Nat Commun 2020;11:1236.
21 Youden WJ. Index for rating diagnostic tests. Cancer 1950;3:32-5.
22 Haugvik SP, Janson ET, Österlund P, et al. Surgical treatment as a principle for patients with high-grade pancreatic neuroendocrine carcinoma: A Nordic multicenter comparative study. Ann Surg Oncol 2016;23:1721-8.
23 Luo Y, Chen X, Chen J, et al. Preoperative prediction of pancreatic neuroendocrine neoplasms grading based on enhanced computed tomography imaging: Validation of deep learning with a convolutional neural network. Neuroendocrinology 2020;110:338-50.
24 Gennatas ED, Wu A, Braunstein SE, et al. Preoperative and postoperative prediction of long-term meningioma outcomes. PLoS One 2018;13:e0204161.
25 Wang L, Dong T, Xin B, et al. Integrative nomogram of CT imaging, clinical, and hematological features for survival prediction of patients with locally advanced non-small cell lung cancer. Eur Radiol 2019;29:2958-67.
26 Kim M, Kang TW, Kim YK, et al. Pancreatic neuroendocrine tumour: Correlation of apparent diffusion coefficient or WHO classification with recurrence-free survival. Eur J Radiol 2016;85:680-7.
27 Jang KM, Kim SH, Lee SJ, et al. The value of gadoxetic acid-enhanced and diffusion-weighted MRI for prediction of grading of pancreatic neuroendocrine tumors. Acta Radiol 2014;55:140-8.
28 Garg A, Tai K. Comparison of statistical and machine learning methods in modelling of data with multicollinearity. Int J Model Identif Control 2013;18:295-312.
29 Zhu X, Thung KH, Adeli E, et al. Maximum Mean Discrepancy Based Multiple Kernel Learning for Incomplete Multimodality Neuroimaging Data. Med Image Comput Comput Assist Interv 2017;10435:72-80.
30 Yang J, Dvornek NC, Zhang F, et al. Unsupervised Domain Adaptation via Disentangled Representations: Application to Cross-Modality Liver Segmentation. Med Image Comput Comput Assist Interv 2019;11765:255-63.
31 Zhang Y, Wei Y, Wu Q, et al. Collaborative unsupervised domain adaptation for medical image diagnosis. IEEE Trans Image Process 2020;29:7834-44.
32 Gao X, Wang X. Deep learning for World Health Organization grades of pancreatic neuroendocrine tumors on contrast-enhanced magnetic resonance images: a preliminary study. Int J Comput Assist Radiol Surg 2019;14:1981-91.
33 Niazi MKK, Tavolara TE, Arole V, et al. Identifying tumor in pancreatic neuroendocrine neoplasms from Ki67 images using transfer learning. PLoS One 2018;13:e0195621.

Cite this article as: Song C, Wang M, Luo Y, Chen J, Peng Z, Wang Y, Zhang H, Li ZP, Shen J, Huang B, Feng ST. Predicting the recurrence risk of pancreatic neuroendocrine neoplasms after radical resection using deep learning radiomics with preoperative computed tomography images. Ann Transl Med 2021;9(10):833. doi: 10.21037/atm-21-25

Supplementary

Table S1 Description of CT findings

CT findings	Subcategories	Description
Primary lesion	The max diameter	The maximal diameter in the axial plane was recorded for pNENs and the data was categorized as ≥20 mm group and <20 mm group
	Location	The location of pNENs was recorded as the uncinate process, head or neck, body and tail
	Property	pNENs was divided as purely solid, purely cystic and solid-cystic mixed types according to the hypo-attenuation portion less than 30 HU with no enhancement in both arterial and portal venous phases
	Calcification	Calcification in pNENs was recorded with the presence of hyper-attenuation portion more than 80 HU
	Shape	The shape of pNENs was classified into 3 types: round shape with clear margin, simple nodular with extra-nodular growth and confluent multinodular
	Boundary	If there was a clear line between pNENs' lesion and surrounding tissues, it was recorded as clear boundary. Otherwise, it was recorded as unclear boundary
	Vessel involvement	If there was filling defect in pNENs' surrounding vessels (artery observed in arterial phase, venous observed in venous phase), it was recorded as surrounding vessel involvement. Otherwise, it was recorded as surrounding vessel non-involvement
	CT ratio	CT ratio was defined as the CT value of pNENs' lesion divided by the non-tumorous pancreatic parenchyma. We recorded CT ratio in unenhanced phase, arterial phase and venous phase, respectively
	Relatively enhanced rate	Relatively enhanced ratio was calculated by that increased CT value of pNENs' lesion divided by the increased CT value of aorta in the same plane. We recorded the data in arterial phase and venous phase, respectively
Pancreas	Pancreatic duct dilated or cut	The dilation of pancreatic duct was recorded when the diameter of main pancreatic duct measured more than 3 mm. Pancreatic duct cut was defined as a sudden interruption of the main pancreatic duct
	Pancreas atrophy	Pancreas atrophy was defined as more than expected loss or of adipose infiltration of pancreas parenchyma
Lymph node	Morphology	The maximal diameter of lymph node short axis in the axial plane was recorded and the data was categorized as normal group (<10 mm), enlarged group (≥10 mm) and multinodular confluent group
	Enhancement pattern	pNENs' lesion was characterized as heterogeneous enhancement when there was hypo-attenuation area in the solid part and homogeneous enhancement when the solid part appeared as the same attenuation in arterial phase
Hepatobiliary system	Fatty liver	Fatty liver was defined as the CT value of liver decreased less than 40 HU
	Focal benign lesion	Hepatic focal benign lesions included pure cyst, focal nodular hyperplasia and calcification with typical CT imaging appearance confirmed by hepatic lesion imaging diagnostic expert
	Bile duct dilatation	The diameter of bile duct >5 mm was recorded as dilation. The diameter of common hepatic duct and common bile duct >10 mm was recorded as dilation
Portal system	Portal vein	The diameter of portal vein was measured at the plane of hepatic hilum
	Splenic vein	The diameter of splenic vein was measured at the plane of splenic hilum
	Splenic varices	Increased and dilated blood vessels at the splenic hilum were recorded as splenic varices
	Splenomegaly	The spleen was beyond 5 costal units on axial plane

CT, computed tomography; pNEN, pancreatic neuroendocrine neoplasm.

http://dx.doi.org/10.21037/atm-21-25

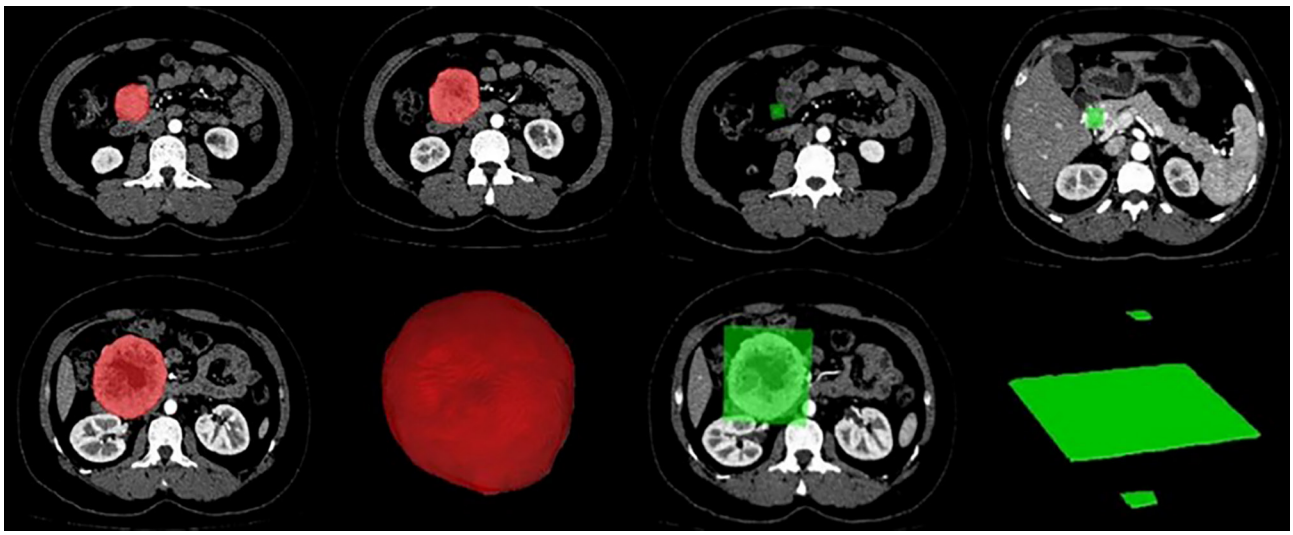

Figure S1 ROIs for radiomics (red) and DLR (green). ROIs, regions of interest; DLR, deep learning radiomics.

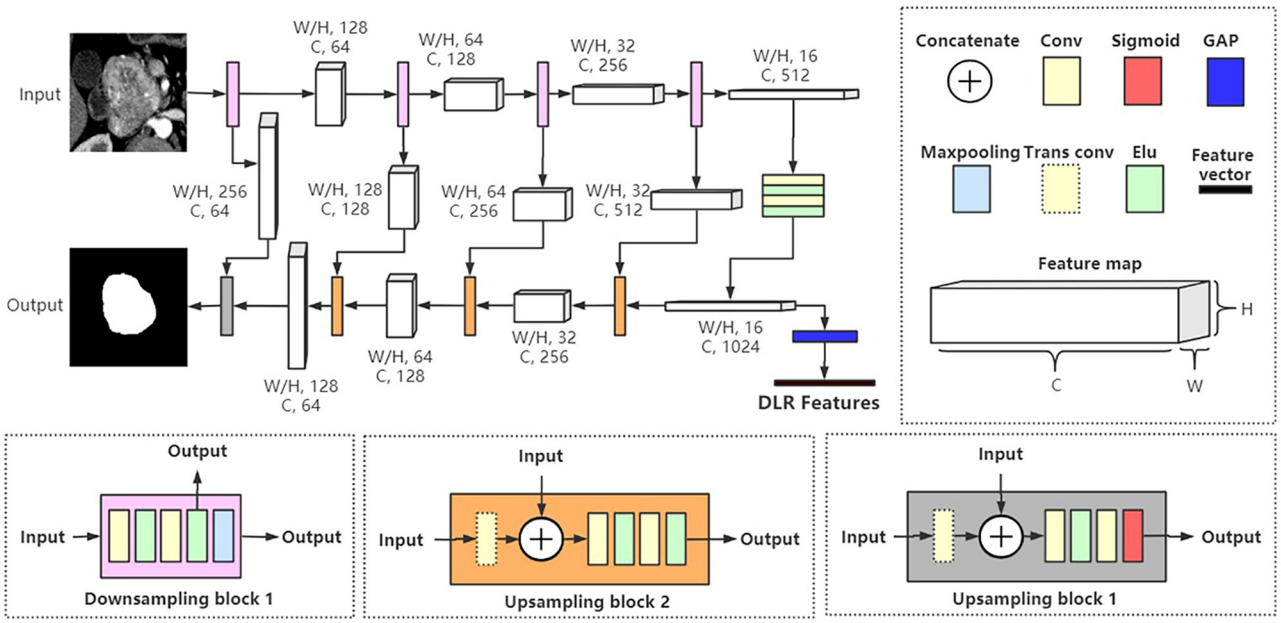

Figure S2 Network structure of 2D U-net. The W and H and C donate width and height and channel of feature map, respectively. Conv, convolution; GAP, global average pooling; Trans conv, transposed convolution.

http://dx.doi.org/10.21037/atm-21-25

Table S2 Clinical information and CT findings in recurrence and recurrence-free pNENs (56 patients from Hospital I)

Variables	Recurrence-free (n=46)	Recurrence (n=10)	Statistics*	P
Clinical information				
Age	14.25 (3.00)	14.75 (4.50)	m 140.000	0.054
Sex			x 0.487	0.730
F	24	4		
M	22	6		
Symptom			f	0.032
N	23	9		
Y	23	1		
Primary lesion				
The max diameter			f	0.032
<20 mm	23	1		
≥20 mm	23	9		
Location			f	0.850
Uncinate process	16	5		
Head and neck	14	2		
Body	4	1		
Tail	12	2		
Property			f	0.054
Cystic	0	1		
Mixed	18	6		
Solid	28	3		
Calcification			f	0.390
Y	38	7		
N	8	3		
Shape			f	0.022
Round	29	2		
Local lobulated	11	4		
Confluent multinodular	6	4		
Boundary			f	0.140
Clear	32	4		
Unclear	14	6		
Vessel involvement				1.000
N	42	9		
Y	4	1		
CT ratio				
Unenhanced	1.16 (0.38)	0.85 (0.45)	m 136.000	0.044
Arterial phase	1.21 (0.61)	0.90 (0.88)	m 172.000	0.260
Venous phase	1.16 (0.38)	0.85 (0.45)	m 136.000	0.044
Relatively enhanced rate				
Arterial phase	0.43 (0.39)	0.28 (0.34)		0.052
Venous phase	0.64 (0.37)	0.48 (0.45)		0.120
Pancreas				
Pancreatic duct dilated or cut			f	0.028
N	39	5		
Y	7	5		
Pancreas atrophy			f	0.680
N	36	7		
Y	10	3		
Lymph node				
Morphology			f	0.023
Normal	41	6		
Enlarged	5	3		
Confluent multinodular	0	1		
Enhancement pattern			f	0.029
Homogeneous	46	8		
Heterogeneous	0	2		
Hepatobiliary system				
Fatty liver			f	1.000
N	43	10		
Y	3	0		
Focal benign lesion			f	0.490
N	26	4		
Y	20	6		
Bile duct dilatation			f	1.000
N	40	9		
Y	6	1		
Portal system				
Portal vein	14.25 (3.00)	14.75 (4.50)	m 189.500	0.380
Splenic vein	8.83±2.15	7.60±2.95	t –1.244	0.240
Splenomegaly			f	1.000
N	29	6		
Y	17	4		
Splenic varices			f	0.560
N	43	9		
Y	3	1		

*, *t* represents Student's *t*-test, m represents Mann Whitney U test, x represent Pearson chi-square test, f represents fisher exact probability test. CT, computed tomography; pNEN, pancreatic neuroendocrine neoplasm.

http://dx.doi.org/10.21037/atm-21-25

Table S3 DeLong test results (P value) of ROC comparisons for all models based on Hospital I image datasets

Model	DLR-A	DLR-V	DLR-A&V	Radiomics-A	Radiomics-V	Radiomics A&V	CT findings
DLR-A	–	0.0632	0.1519	0.5952	0.2808	0.4309	0.1191
DLR-V	–	–	0.1618	0.1719	0.3590	0.2756	0.7364
DLR-A&V	–	–	–	0.8552	0.6310	0.9041	0.2474
Radiomics-A	–	–	–	–	0.6500	0.7994	0.1046
Radiomics-V	–	–	–	–	–	0.5058	0.2855
Radiomics-A&V	–	–	–	–	–	–	0.1966
CT findings	–	–	–	–	–	–	–

ROC, receiver operating characteristic; DLR, deep learning radiomics; A, arterial; V, venous; A&V, arterial & venous; CT, computed tomography.

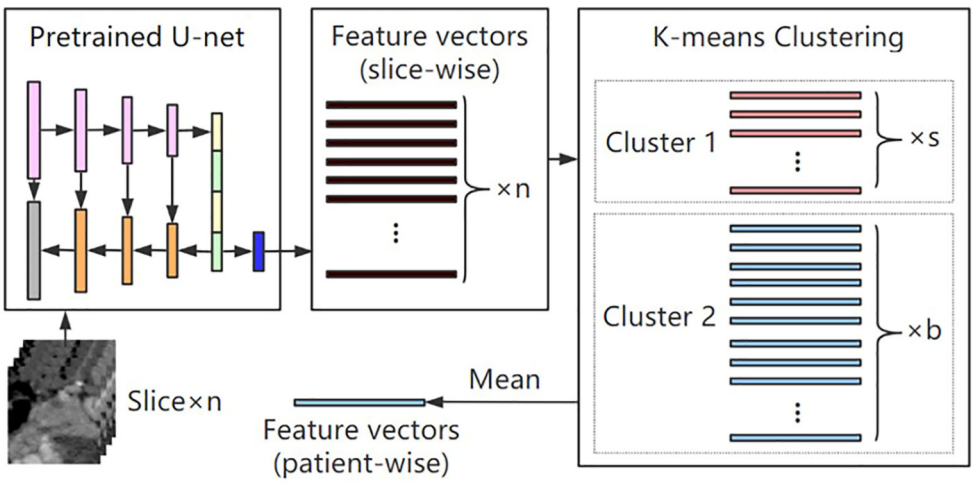

Figure S3 Flow chart of deep learning radiomics feature extraction.

http://dx.doi.org/10.21037/atm-21-25

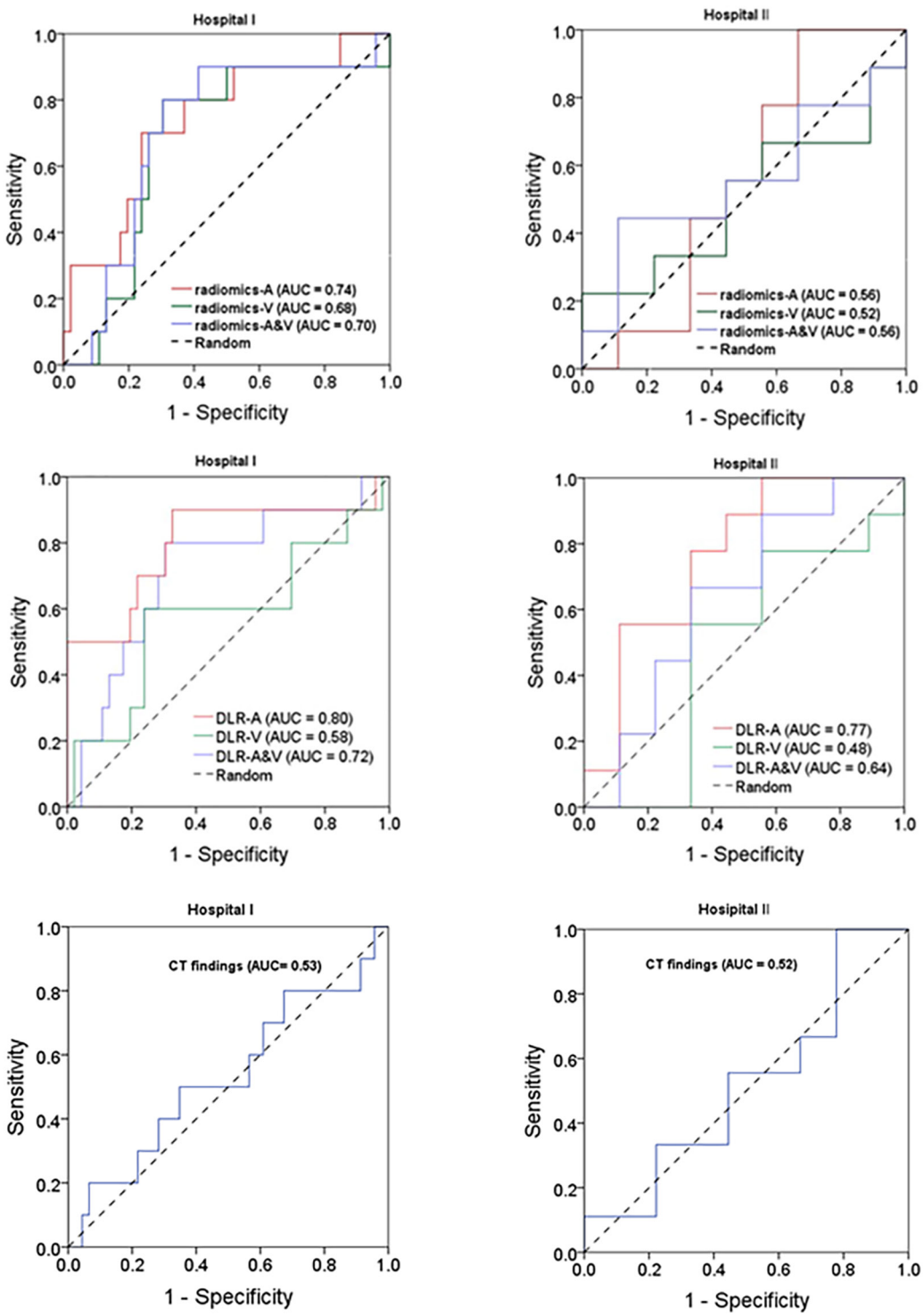

Figure S4 ROCs of different phases with radiomics, deep learning radiomics (DLR) and CT findings in the internal and external groups. ROC, receiver operating characteristic.

http://dx.doi.org/10.21037/atm-21-25

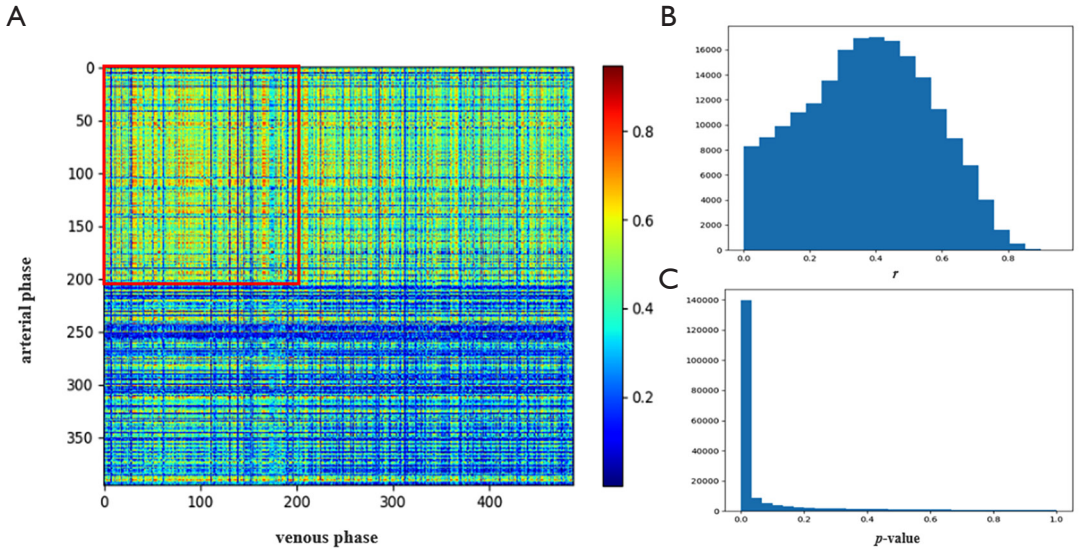

Figure S5 Collinearity analysis results of DLR features in arterial and venous phases. (A) Heatmap of absolute value of correlation coefficient r, 206 features overlap between arterial and venous phases (which correlation coefficients are shown in the red box). (B) Histogram of Pearson's correlation coefficient r distribution. (C) Histogram of P value distribution. DLR, deep learning radiomics.

Supplementary Materials and Methods

Section 1 Training U-net for DLR

We used a 2D U-net to extract DLR features (*Figure S2*). The encoder of the U-net contained 4 downsampling modules, and the decoder contained 4 upsampling modules constructed based on transposed convolution. Skip connections were set between the upsampling and downsampling modules to provide more high-resolution information for the decoder. The initial learning rate was set as 1×10^{-5}, the optimizer was Adam, and we used cross-entropy as loss function. Dice similarity coefficient (DSC) was calculated on the validation set for evaluating the performance of segmentation, and the calculation formula of DSC was as follow, where A and B are the ground truth (GT) and predicted segmentation mask of the image, respectively.

$$\mathrm{DSC}(A,B)=\frac{2|A\cap B|}{|A|+|B|} \quad [1]$$

Section 2 DLR features extraction

In the feature extraction process (*Figure S3*), we first took the smallest externalized cube of the region of interest (ROI) roughly annotated by the radiologists in 3D space as processed ROI, then for each patient we inputted each slice of CT image in processed ROI and extracted the feature map [after exponential linear unit (ELU) activation] of the last convolution layer before the decoder. Then a global average pooling (GAP) was performed to convert the feature map with size of 16×16×1,024 into a feature vector with size of 1×1,024.

The input of segmentation network was a 2D slice of the tumor on CT image, and the recurrence annotation was patient-wise, so it was necessary to aggregate all slice-wise feature vectors of the same patient into a patient-wise feature vector. The feature vectors extracted from the multi-layer images of the same sample was n ×1,024, and n was the number of tumor slices. All feature vectors were clustered into 2 clusters based on K-means algorithm, and the maximum cluster was preserved. Then we took the mean value in the maximum cluster along feature dimension to get the final vector with a size of 1×1,024.

Section 3 Model integration

For model integration, we used models in each fold of cross-validation on internal group to predict the recurrence risk of each patient in external group, and the average of the multi-model predicted recurrence risk was used to calculate the evaluation metric. The whole process of model integration can be expressed as following equation,

$$Y_i=\left\{F(x_{i,p})\mid x_{i,p}\in X_i\right\} \quad [2]$$

$$Z_i=\frac{1}{N}\sum\nolimits_n g_n(K(Y_i)) \quad [3]$$

where X and x represent the CT image (in processed ROI) and its slice, respectively. And i is the patient index, p is the slice index. F denotes the segmentation feature extraction process (whose output is a feature vector), and Y is the feature vector set of all slices of tumor X. In the latter formula, K is the feature aggregation operation (K-means clustering), and g denotes the recurrence prediction model (whose input is a feature vector). N is the number of classification models, and n is the cross-validation model index. Z is the final predicted recurrence risk of patient in external group.

http://dx.doi.org/10.21037/atm-21-25