A Deep Transfer Learning-Based Radiomics Model for Prediction of Local Recurrence in Laryngeal Cancer

Y Jia[1,2], X. Sharon Qi[2], J Du[2], R Chin[2], M Elizabeth[2], K Sheng[2,*]

1. Shaanxi Key Laboratory of Network Data Intelligent Processing, Xi'an University of Posts and Communications，Xi'an, China

2. Department of Radiation Oncology, University of California, Los Angeles, CA

JULY 12–16 2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

INTRODUCTION

Local recurrence (LR) is one of the dominant forms of treatment failure for patients with advanced head and neck (H&N) cancer and 13–35% of patients will suffer from LR. Laryngeal cancer is the second most common type of H&N cancer. Early prediction of LR for the patients will help develop an individualized treatment to minimize the risk and reliable prognostic biomarkers are needed. In several radiomics prediction researches[1-3], the AUCs of LR prediction were always the lowest on different patient cohorts when compared with other endpoints, such as DM (distant metastases) and OS (overall survival). So we took a closer look at the LR prediction in this study. The average age of the cases of H&N cancer is 65 and nearly half of the CT scans have artifacts caused by metal dental implants in oropharynx[4]. The artifacts can induce big changes in the commonly used texture features and worsen the performance of radiomics modeling. According to these published results of LR prediction and the studies of metal streak artifacts, the prediction performance appears to be limited compared to the acceptance criterion for eligibility (0.9). We did some deeper analysis of all the steps including image analysis of the original CT, delineation of the GTV contour, CNN model design and decision strategy. Based on the analyzation result, we built a 2.5D transfer-learning model (using the 2D slices for training and 3D images for decision-making) for LR prediction of laryngeal cancer and obtained an AUC higher than 0.91.

AIM

Most of the existing radiomics research of the head and neck (H&N) cancer extracts quantitative features as prognostic factors to predict the clinical endpoints. However, the prediction performance of local recurrence (LR) is still challenging. We aimed to develop and validate a 2.5 dimensional (2.5D) transfer learning-based DCNN (Deep Convolutional Neural Network) model for local recurrence (LR) prediction in laryngeal cancer.

METHOD

A total of 48 laryngeal cancer patients with pre-treatment Computed Tomography (CT) from The Cancer Imaging Archive (TCIA) were included in this study (LR=37, No LR=11). 403 2D patches derived from the gross tumor volume (GTV) on the planning CT were extracted and refined as the dataset for model training. Figure 1 is the workflow of this research. All GTV contours are provided for the CT scans based on PET images. Firstly, we extracted the primary site with the GTV contours and refined the tumor region with image processing methods. Secondly, we designed an LR and non-LR classification model transferred from a CNN model pre-trained with ImageNet and fine-tuned the model with our training data. Finally, we tested the model with 5-folds cross-validation to evaluate the performance of model. Details of the methods are described below.

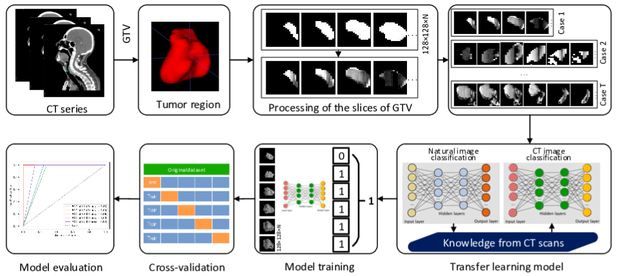

Figure 1 LR risk assessment workflow. The upper side of the workflow is processing of CT images. The lower part is the design of the prediction model.

1. In previous experiments, we found that some of the GTV contours of the tumor located in larynx have bones inside, while bones present high signal intensity in the image and it becomes the noise for tumor parenchyma analysis. We defined the laryngeal parenchyma as the homogeneous tissue region excluding the bone. We also used z-score normalization and prepared images for the model training.
2. Transfer a CNN model from VGG19 pre-trained on ImageNet, and fine-tune on the tumor patches.
3. Use image augmentation and weighted loss function to address the data imbalance.
4. Combine a stack of 2D slices through a majority vote to provide the final prediction for a given patient.

RESULTS

1. Because most of the studies of H&N cancer outcome prediction includes tumors in the larynx, and other H&N cancer regions, such as nasopharynx, oropharynx and hypopharynx, it is hard to compare the proposed model with other state-of-the-art impartially. However, it is also crucial to show that the prediction model for laryngeal cancer works much better than the model mixing all types of cancers in H&N together. Some of the related LR prediction metrics from related references are shown in Table 1.
2. We tested the model with a different type of input images, the segmented ROI within GTV and the BoundingBox (BBOX). The results for using refined GTV as input were better than using the BBOX for LR prediction, with the AUC of 0.91/0.64, the sensitivity of 1/0.69, the specificity of 0.81/0.59 for refined GTV and BBOX, respectively, as shown in Figure 2. Since the model with BBOX did not perform well for 2.5D classification, we use refined GTV in this study.
3. Figure 3 is the Grad-CAM (Gradient-weighted Class Activation Mapping) of the tumor. Red regions correspond to high scores for the class of LR. It shows us that the surrounding parenchyma in the upper right and upper left corners (higher than the removed bones) is not salient in the Grad-CAM. We can use the information to propose new ROI refine strategies to constrain the input and optimize the prediction model potentially.

Table 1 Comparison of LR prediction performance of models constructed for laryngeal cancer with other combinations of variables in H&N cancer.

Task	Image modality/Input variables	Prediction			Method
		AUC	Sensitivity	Specificity	
LR prediction in H&N cancer (including tumors in both oropharynx and larynx) (N=137, 8-folds cross-validation)	Clinical	0.65±0.02	0.21±0.03	0.87±0.03	handcrafted features +random forest
	CT	0.57±0.02	0.13±0.02	0.95	
	CT+Clinical	0.62±0.01	0.07±0.02	0.96	
	CT+Clinical, all slices	0.48±0.03	0.17±0.08	0.80±0.01	2.5D CNN transferred from VGG19
	CT, all slices	0.53±0.04	0.19±0.07	0.87±0.01	
LR prediction in laryngeal cancer (N=48, 8-folds cross-validation)	Clinical	0.72±0.03	0.54±0.07	0.90±0.01	handcrafted features +random forest
	CT	0.63±0.07	0.38±0.14	0.88±0.01	
	CT+Clinical	0.63±0.05	0.36±0.08	0.90±0.03	
	CT+Clinical, all slices	0.90±0.02	1.00	0.80±0.03	2.5D CNN transferred from VGG19
	CT, all slices	0.91±0.02	1.00	0.79±0.04	

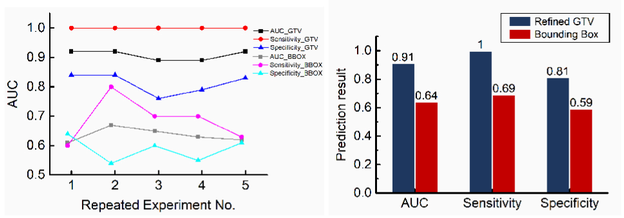

Figure 2 Comparison of the prediction result with different input images. Testing result of LR and no-LR prediction of laryngeal cancer with slices of ROI and slices of BBOX. (a) Results of the five parallel experiments. (b) Average value of the five parallel experiments.

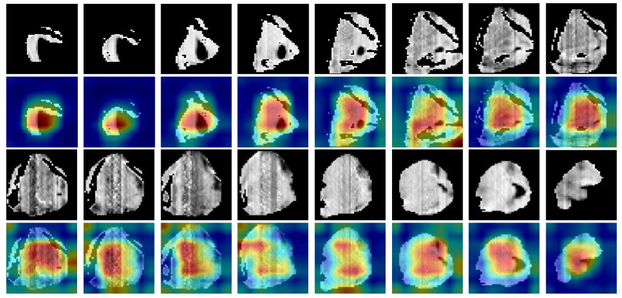

Figure 3 Normalized input tumor patches and the Grad-CAM visualizations for the last convolutional layer of our model. Patches are extracted from the same tumor.

DISCUSSION

We admit there are still some limitations of this study. One is the defined GTV of the primary site. We found that in this dataset from TCIA, more than 38% of the patients had dental implant artifacts, and the primary site and lymph nodes were not separated. We also found the way of GTV delineation varies differently, such as including the bone or not, the location of the top and bottom slice along the z-axis. It is necessary to refine the original GTV to enhance the accuracy of tumor extraction. Another limitation is the dataset imbalance and lacking enough multi-institutional dataset. We need to collect more data to validate the performance of deep learning algorithm.

CONCLUSIONS

This study demonstrated that some factors of the tumor regions, such as the definition of tumor regions (GTV or BBOX), the dental implant artifacts of the CT scans, and the post-processing methods of GTV are all associated with the performance of the prediction model. By eliminating these factors, we built a 2.5D LR prediction model for laryngeal cancer, which performed much better when compared to previous studies mixing different types of cancers in H&N together. These results indicate that it is important to look deeper in prognostic imaging markers for LR prediction in H&N cancer. Accurately extracted tumor regions and imaging features may ultimately promote the prognostic power of radiomics model. To our knowledge, this is the first study that shows deep learning algorithm works much better in laryngeal cancer than in the H&N cancers including tumors in both oropharynx and larynx.

ACKNOWLEDGEMENTS

This research is funded by the Key Research and Development Program of Shaanxi Province (2019GY-021), the Foundation of Shaanxi Educational Committee (18JK0722), and CSC Scholarship. The authors gratefully acknowledge all these supports.

REFERENCES

1 **Diamant A, Chatterjee A, Vallières M, Shenouda G, Seuntjens J.** Deep learning in head & neck cancer outcome prediction. Scientific reports. *2019;9(1):2764.*

2 **Vallières M, Kay-Rivest E, Perrin LJ, et al.** Radiomics strategies for risk assessment of tumour failure in head-and-neck cancer. Scientific reports. *2017;7(1):10117.*

3 **Zhai T-T, Langendijk JA, van Dijk LV, et al.** The prognostic value of CT-based image-biomarkers for head and neck cancer patients treated with definitive (chemo-) radiation. Oral Oncology. *2019;95:178-186.*

4 **Wee L, Dekker A.** Data from Head-Neck-Radiomics-HN1. In: Archive TCI, ed.: The Cancer Imaging Archive; *2019.*

CONTACT INFORMATION

Ke Sheng, Email: KSheng@mednet.ucla.edu
Yang Jia, Email: jiayang@xupt.edu.cn