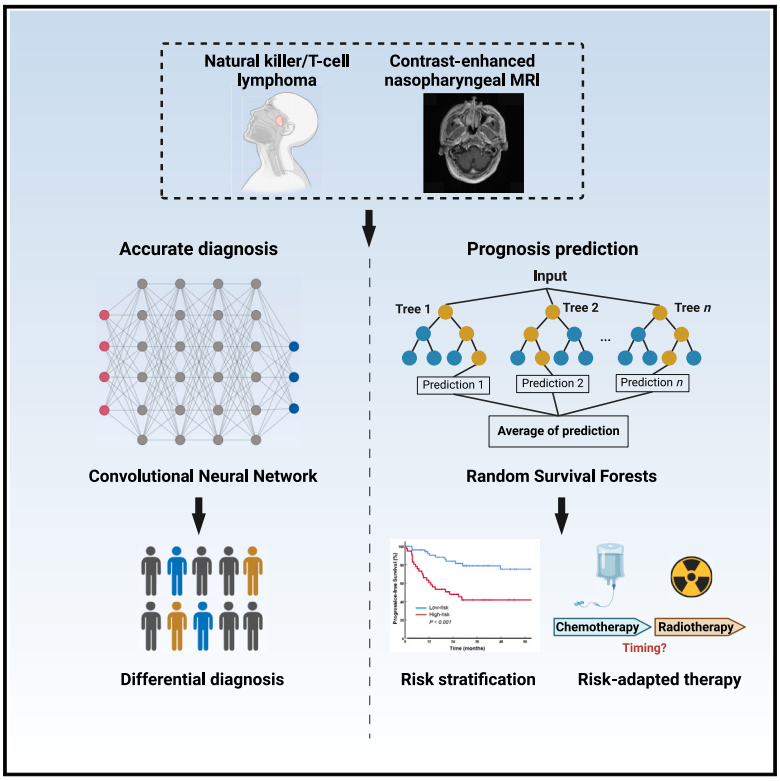
Cell Reports
Medicine

Article

Artificial intelligence for diagnosis and prognosis prediction of natural killer/T cell lymphoma using magnetic resonance imaging

Graphical abstract



Highlights

- AI systems provide additional advantages for MRI image evaluation
- Diagnostic system contributes to differential diagnosis of nasopharyngeal lesions
- Prognostic system predicts prognosis and may assist therapeutic optimization for NKTCL

Authors

YuChen Zhang, YiShu Deng, QiHua Zou, ..., Liang Wang, ChaoFeng Li, QingQing Cai

Correspondence

xchuanm@sysucc.org.cn (C.X.), wangliangtrhos@126.com (L.W.), lichaofeng@sysucc.org.cn (C.L.), caiqq@sysucc.org.cn (Q.C.)

In brief

Zhang et al. construct and validate predictive systems using artificial intelligence algorithms and magnetic resonance imaging that are capable of assisting with accurate diagnosis and prognosis prediction of natural killer/T cell lymphoma and may give implications for therapeutic preferences of early-stage patients.

Zhang et al., 2024, Cell Reports Medicine 5, 101551
May 21, 2024
https://doi.org/10.1016/j.xcrm.2024.101551

CellPress

Cell Reports Medicine

CellPress
OPEN ACCESS

Article

Artificial intelligence for diagnosis and prognosis prediction of natural killer/T cell lymphoma using magnetic resonance imaging

YuChen Zhang,[1,2,19] YiShu Deng,[1,3,4,19] QiHua Zou,[1,2,19] BingZhong Jing,[1,3,19] PeiQiang Cai,[1,5,19] XiaoPeng Tian,[1,2,19] Yu Yang,[6] BingZong Li,[7] Fang Liu,[8] ZhiHua Li,[9] ZaiYi Liu,[10,11] ShiTing Feng,[12] TingSheng Peng,[13] YuJun Dong,[14] XinYan Wang,[15] GuangYing Ruan,[1,5] Yun He,[1,5] ChunYan Cui,[1,5] Jiao Li,[1,5] Xiao Luo,[1,5] HuiQiang Huang,[1,2] HaoHua Chen,[1,3] SongQi Li,[16] Ying Sun,[1,17] ChuanMiao Xie,[1,5,*] Liang Wang,[18,*] ChaoFeng Li,[1,3,*] and QingQing Cai[1,2,20,*]

[1]State Key Laboratory of Oncology in South China, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangdong Provincial Clinical Research Center for Cancer, Sun Yat-sen University Cancer Center, Guangzhou 510060, P.R. China
[2]Department of Medical Oncology, Sun Yat-sen University Cancer Center, Guangzhou, P.R. China
[3]Information Technology Center, Sun Yat-sen University Cancer Center, Guangzhou 510060, P.R. China
[4]School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, P.R. China
[5]Department of Radiology, Sun Yat-Sen University Cancer Center, Guangzhou, P.R. China
[6]Department of Lymphadenoma and Head & Neck Medical Oncology, Fujian Provincial Cancer Hospital & Institute, Fuzhou, P.R. China
[7]Department of Hematology, The Second Affiliated Hospital of Suzhou University, Jiangsu, P.R. China
[8]Department of Pathology, The First People's Hospital of Foshan, Foshan, P.R. China
[9]Department of Oncology, Sun Yat-sen Memorial Hospital, Guangzhou, Guangdong, P.R. China
[10]Department of Radiology, Guangdong Provincial People's Hospital (Guangdong Academy of Medical Sciences), Southern Medical University, Guangzhou 510080, P.R. China
[11]Guangdong Provincial Key Laboratory of Artificial Intelligence in Medical Image Analysis and Application, Guangzhou 510080, P.R. China
[12]Department of Radiology, The First Affiliated Hospital, Sun Yat-Sen University, Guangzhou 510080, P.R. China
[13]Department of Pathology, The First Affiliated Hospital of Sun Yat-sen University, Guangzhou 510080, P.R. China
[14]Department of Hematology, Peking University First Hospital, Beijing 100034, P.R. China
[15]Department of Radiology, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, P.R. China
[16]Zhongshan School of Medicine, Sun Yat-sen University, Guangzhou 510080, P.R. China
[17]Department of Radiation Oncology, Sun Yat-sen University Cancer Center, Guangzhou, P.R. China
[18]Department of Hematology, Beijing Tongren Hospital, Capital Medical University, Beijing 100730, P.R. China
[19]These authors contributed equally
[20]Lead contact
*Correspondence: xchuanm@sysucc.org.cn (C.X.), wangliangtrhos@126.com (L.W.), lichaofeng@sysucc.org.cn (C.L.), caiqq@sysucc.org.cn (Q.C.)
https://doi.org/10.1016/j.xcrm.2024.101551

SUMMARY

Accurate diagnosis and prognosis prediction are conducive to early intervention and improvement of medical care for natural killer/T cell lymphoma (NKTCL). Artificial intelligence (AI)-based systems are developed based on nasopharynx magnetic resonance imaging. The diagnostic systems achieve areas under the curve of 0.905–0.960 in detecting malignant nasopharyngeal lesions and distinguishing NKTCL from nasopharyngeal carcinoma in independent validation datasets. In comparison to human radiologists, the diagnostic systems show higher accuracies than resident radiologists and comparable ones to senior radiologists. The prognostic system shows promising performance in predicting survival outcomes of NKTCL and outperforms several clinical models. For patients with early-stage NKTCL, only the high-risk group benefits from early radiotherapy (hazard ratio = 0.414 vs. late radiotherapy; 95% confidence interval, 0.190–0.900, p = 0.022), while progression-free survival does not differ in the low-risk group. In conclusion, AI-based systems show potential in assisting accurate diagnosis and prognosis prediction and may contribute to therapeutic optimization for NKTCL.

INTRODUCTION

Natural killer/T cell lymphoma (NKTCL) is a unique subtype of non-Hodgkin's lymphoma, with the most common involvement site of the upper aerodigestive tract, such as the nasopharynx, leading to its insidious onset.[1,2] Nasopharyngeal carcinoma (NPC) and NKTCL have significantly higher incidence in East Asia and are the two top ranked types of malignancies that occur in nasopharynx.[3–5] Clinical manifestations of NKTCL in early stages are similar to NPC, but prognosis and treatment options

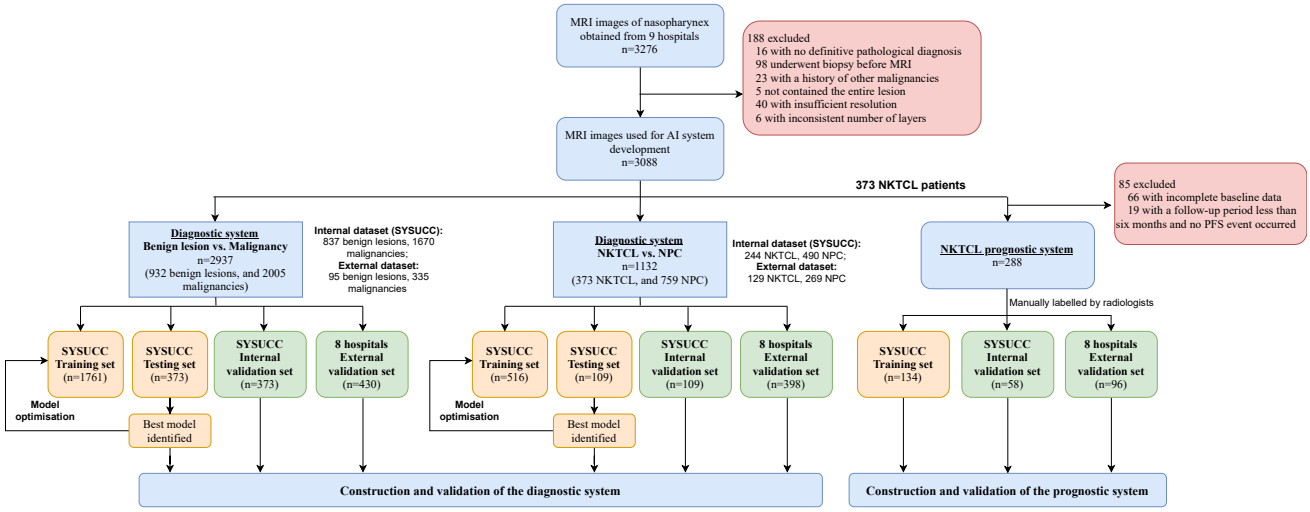


Figure 1. The workflow of construction and validation of the diagnostic and prognostic system

AI, artificial intelligence; MRI, magnetic resonance imaging; NKTCL, natural killer/T cell lymphoma; SYSUCC, Sun Yat-sen University Cancer Center.

are distinct.[5,6] Detection of Epstein-Barr virus (EBV) infection is not helpful in differential diagnosis since both NKTCL and NPC are associated with EBV infection.[4,5] Timely and accurate diagnosis of nasopharyngeal malignancies may contribute to early intervention and prognosis improvement.

Magnetic resonance imaging (MRI) of the nasopharynx is a standard-of-care imaging modality at the initial diagnosis of a nasopharyngeal tumor.[7] Given its high resolution of soft tissue, MRI provides more detailed characteristics of the local lesions, such as signal asymmetry, focal loss of deep mucosal, and extension of the tumor; MRI is also important in identifying occult lesions in the nasopharynx.[7–9] However, minute differences in MRI images are often subtle and difficult to discern with human eyes, which drives us to use a new technique to extract that information.

Artificial intelligence (AI) can automatically identify and analyze features at the pixel level and has been increasingly used to aid in clinical decision-making in a variety of medical scenarios.[10,11] Deep learning approaches have been demonstrated to enhance the accuracy and efficiency of imaging diagnosis, including but not limited to radiological, endoscopic, and pathological images.[12–15] AI approaches have also been used to predict prognosis in patients with lymphoma.[16–18] To our knowledge, currently, there is no previous report of a diagnostic or prognostic model for NKTCL based on MRI and AI algorithms.

In this study, we constructed a two-step AI diagnostic system to detect nasopharyngeal malignant lesions, and distinguish NKTCL vs. NPC, by using MRI data from 9 hospitals in China. Furthermore, we developed an AI prognostic system to predict progression-free survival (PFS) in patients with NKTCL.

RESULTS

The workflow of this study is shown in Figure 1. A total of 3,276 patients who met the eligibility criteria were screened, and finally 3,088 cases were enrolled for further development of diagnostic systems. Patients from Sun Yat-sen University Cancer Center (SYSUCC) were used for training, testing, and internal validation, while patients from eight other participating hospitals were used for external validation. Baseline characteristics of the patients from each dataset for the development of the diagnostic system are shown in Table 1. The age of the included patients was between 18 and 83 years. The percentages of male patients ranged from 62.5% to 72.1% in the cohort for discrimination of benign and malignant lesions and from 71.5% to 74.1% in the cohort for differential diagnosis of NKTCL and NPC. Representative images of different histological types of nasopharyngeal lesions are shown in Figure S1.

The development of the prognostic system was conducted using 288 patients with NKTCL with complete baseline clinical data. Patients from SYSUCC were used for training and internal validation, while patients from eight other centers were used for external validation. Baseline characteristics of the patients with NKTCL for the development of prognostic system are shown in Table S1.

Model performance of the diagnostic system

A deep-learning-based 3D DenseNet model was used to construct the diagnostic systems. The training curves exhibited similar data loss in the training and testing sets, indicating sufficient robustness (Figure S2). The model accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) in the internal and external validation sets are summarized in Table 2.

In the task of detecting nasopharyngeal malignant lesions, the diagnostic system showed areas under the curve (AUCs) of 0.960 (95% confidence interval [CI], 0.935–0.977) and 0.908 (95% CI, 0.879–0.931) in an internal validation set and an independent external validation set, respectively (Figure 2A). The accuracy, sensitivity, and specificity were 90.6% (95% CI, 87.2%–93.4%), 0.919 (95% CI, 0.878–0.950), and 0.872 (95% CI, 0.800–0.925) in internal validation and 85.9% (95% CI, 82.6%–88.8%), 0.863 (95% CI, 0.826–0.894), and 0.842 (95% CI, 0.753–0.909) in external validation.

Table 1. Baseline participants' characteristics

	Benign vs. malignant lesions				NPC vs. NKTCL			
	Training	Testing	Internal validation	External validation	Training	Testing	Internal validation	External validation
Sample size	1,761	373	373	430	516	109	109	398
Age								
Median (range)	46 (18–83)	46 (20–77)	47 (18–74)	48 (18–80)	45 (18–82)	48 (20–76)	44 (19–74)	47 (18–78)
≤60 years (%)	1,562 (88.7)	331 (88.7)	323 (86.6)	356 (82.8)	465 (90.1)	94 (86.2)	103 (94.5)	341 (85.7)
>60 years (%)	199 (11.3)	42 (11.3)	50 (13.4)	74 (17.2)	51 (9.9)	15 (13.8)	6 (5.5)	57 (14.3)
Gender (%)								
Male	1,157 (65.7)	233 (62.5)	264 (70.8)	310 (72.1)	369 (71.5)	79 (72.5)	79 (72.5)	295 (74.1)
Female	604 (34.3)	140 (37.5)	109 (29.2)	120 (27.9)	147 (28.5)	30 (27.5)	30 (27.5)	103 (25.9)
Diseases (%)								
NKTCL	123 (7.0)	25 (6.7)	25 (6.7%)	49 (11.4)	172 (33.3)	36 (33.0)	36 (33.0)	129 (32.4)
NPC	968 (55.0)	207 (55.5)	207 (55.5%)	269 (62.6)	344 (66.7)	73 (67.0)	73 (67.0)	269 (67.6)
Other lymphomas[a]	38 (2.2)	7 (1.9)	7 (1.9)	7 (1.6)	–	–	–	–
Other malignancies[b]	45 (2.6)	9 (2.4)	9 (2.4)	95 (22.1)	–	–	–	–
Benign lesions[c]	587 (33.3)	125 (33.5)	125 (33.5)	269 (62.6)	–	–	–	–

Data are shown as number (%) or median (range). The sum of some percentages may not equal 100% because of rounding. NKTCL, natural killer/T-cell lymphoma; NPC, nasopharyngeal carcinoma.

[a]Other lymphomas include diffuse large B cell lymphoma, follicular lymphoma, high grade B cell lymphoma, extranodal marginal zone lymphoma of mucosa-associated lymphoid tissue, mantle cell lymphoma, peripheral T cell lymphomas - not otherwise specified, angioimmunoblastic T cell lymphoma, and T-lymphoblastic lymphoma.

[b]Other malignancies include adenocarcinoma, salivary glandular tumor, neuroendocrine carcinoma, sarcoma, plasmacytoma, and melanoma.

[c]Benign lesions include mucosa chronic inflammatory, adenoids/lymphoid hyperplasia, atypical hyperplasia, nasopharyngeal cyst, meningioma, necrotic tissue, fungal infection, tuberculosis, and a combination of multiple histopathological features.

In the task of discrimination of NKTCL and NPC, the AUCs were 0.932 (95% CI, 0.867–0.971) and 0.905 (95% CI, 0.872–0.932) in the internal and external validation sets, respectively (Figure 2B). The accuracy, sensitivity, and specificity were 87.2% (95% CI, 79.4%–92.8%), 0.889 (95% CI, 0.739–0.969), and 0.863 (95% CI, 0.762–0.932) in internal validation and 84.2% (95% CI, 80.2%–87.6%), 0.865 (95% CI, 0.793–0.919), and 0.835 (95% CI, 0.785–0.877) in external validation. The corresponding confusion tables in the internal and external validation sets are shown in Figure S3.

Subgroup analyses were performed in internal and external validation sets to evaluate the performance of the AI system across different severity levels of cases. The results indicated stable performance of the model across different malignancy subtypes (NPC/NKTCL/others) and between different radiological characteristics (mucosal thickening/mass) in differential diagnosis of benign vs. malignant lesions (Figures S4A–S4D). The model also performed well between NKTCL cases with or without extensive primary tumors in the discrimination of NKTCL vs. NPC (Figures S4E and S4F). The diagnostic system showed similar performances between different magnetic field strengths (3.0/1.5 T) in both of the diagnostic tasks (Figures S5A–S5D).

Comparison of the diagnostic system with radiologists

Two subsets of 200 cases each (subset A: 48 benign nasopharyngeal lesions and 152 malignancies; subset B: 57 NKTCL and 143 NPC) from the internal and external validation sets were randomly selected to compare the diagnostic capacity of the AI system with radiologists with different degrees of seniorities (2 residents, 2 attending doctors, and 2 senior doctors in radiological diagnosis). The AUCs of the diagnostic system were 0.920 (95% CI, 0.874–0.954) for detecting malignant lesions and 0.914 (95% CI, 0.866–0.949) for discriminating NKTCL from NPC (Figures 2C and 2D). The accuracy, sensitivity, specificity, PPV, and NPV are shown in Table 2.

For detecting malignancies, the diagnostic accuracy of the AI system was 0.875 (95% CI, 0.821–0.917), which was higher than the resident radiologists (0.720 [95% CI, 0.673–0.763], $p < 0.001$) and comparable to the attending radiologists (0.818 [95% CI, 0.776–0.854], $p = 0.072$) and the senior radiologists (0.888 [95% CI, 0.852–0.917], $p = 0.653$). The PPV of the AI system was 0.927 (95% CI, 0.873–0.965), which was similar to all the three radiologists ($p > 0.05$). Meanwhile, the NPV of the AI system (0.744 [95% CI, 0.621–0.853]) was significantly higher than the resident radiologists (0.496 [95% CI, 0.417–0.572], $p = 0.001$) and similar to the attending radiologists (0.632 [95% CI, 0.544–0.711], $p = 0.101$) and the senior radiologists (0.737 [95% CI, 0.658–0.810], $p = 0.872$).

For distinguishing NKTCL from NPC, the accuracy of the AI system (0.855 [95% CI, 0.798–0.901]) was significantly higher than the resident radiologists (0.620 [95% CI, 0.570–0.668], $p < 0.001$) and the attending radiologists (0.760 [95% CI, 0.715–0.801], $p = 0.007$) and comparable to the senior radiologists (0.858 [95% CI, 0.819–0.890], $p = 0.934$). The PPV and NPV of the AI system were 0.464 (95% CI, 0.322–0.605) and 0.981 (95% CI, 0.942–0.996), respectively, which were higher than the residents (PPV: 0.199 [95% CI, 0.142–0.272], $p < 0.001$; NPV: 0.908 [95% CI, 0.865–0.942], $p = 0.006$). Similar

Table 2. Performance of the diagnostic system in internal and external validation sets

Dataset	Accuracy (95% CI)	Sensitivity (95% CI)	Specificity (95% CI)	PPV (95% CI)[a]	NPV (95% CI)[a]
Internal validation set					
Benign vs. malignant	0.906 (0.872–0.934)	0.919 (0.878–0.950)	0.872 (0.800–0.925)	0.934 (0.900–0.958)	0.845 (0.781–0.893)
NPC vs. NKTCL[b]	0.872 (0.794–0.928)	0.889 (0.739–0.969)	0.863 (0.762–0.932)	0.762 (0.640–0.852)	0.940 (0.862–0.976)
External validation set					
Benign vs. malignant	0.865 (0.829–0.896)	0.872 (0.831–0.906)	0.842 (0.753–0.909)	0.951 (0.924–0.969)	0.650 (0.581–0.714)
NPC vs. NKTCL	0.842 (0.802–0.876)	0.865 (0.793–0.919)	0.835 (0.785–0.877)	0.482 (0.415–0.550)	0.971 (0.956–0.981)
AI vs. radiologists[c]					
Benign vs. malignant					
AI system	0.875 (0.821–0.917)	0.895 (0.835–0.939)	0.813 (0.674–0.911)	0.927 (0.873–0.965)	0.744 (0.621–0.853)
Resident	0.720 (0.673–0.763)	0.704 (0.649–0.755)	0.771 (0.674–0.850)	0.891 (0.843–0.928)	0.496 (0.417–0.572)
Attending	0.818 (0.776–0.854)	0.826 (0.778–0.867)	0.792 (0.697–0.868)	0.913 (0.872–0.944)	0.632 (0.544–0.711)
Senior	0.888 (0.852–0.917)	0.875 (0.832–0.910)	0.927 (0.856–0.970)	0.970 (0.941–0.987)	1.737 (0.658–0.810)
NPC vs. NKTCL					
AI system	0.855 (0.798–0.901)	0.895 (0.785–0.969)	0.839 (0.769–0.895)	0.464 (0.322–0.605)	0.981 (0.942–0.996)
Resident	0.620 (0.570–0.668)	0.588 (0.492–0.679)	0.633 (0.574–0.689)	0.199 (0.142–0.272)	0.908 (0.865–0.942)
Attending	0.760 (0.715–0.801)	0.763 (0.674–0.838)	0.759 (0.705–0.807)	0.330 (0.247–0.418)	0.954 (0.921–0.975)
Senior	0.858 (0.819–0.890)	0.825 (0.742–0.889)	0.871 (0.826–0.907)	0.498 (0.387–0.602)	0.970 (0.946–0.987)

AI, artificial intelligence; CI, confidence interval; NKTCL, natural killer/T cell lymphoma; NPC, nasopharyngeal carcinoma; NPV, negative predictive value; PPV, positive predictive value.
[a]PPV and NPV were calculated based on disease proportion in patients included between 2019 and 2021 in the corresponding dataset.
[b]For presentation, "positive" and "negative" refer to NKTCL and NPC, respectively.
[c]Each dataset contained 200 randomly selected cases from internal and external validation sets.

levels of PPV and NPV were observed between AI and attending/senior radiologists.

Model performance of the NKTCL prognostic system

A random survival forest model was used to develop the prognostic systems based on radiomic features (MRI score) or radiomic features and clinical variables (total score) for NKTCL. The median follow-ups were 56.2 (95% CI, 46.0–66.4), 42.1 (95% CI, 30.5–53.6), and 41.0 (95% CI, 26.8–55.1) months for training, internal validation, and external validation sets, respectively. The 3-year time-dependent AUCs of the total score, MRI score, and several established clinical models (the international prognostic index [IPI], Korean prognostic index [KPI], and prognostic index of natural killer lymphoma [PINK]) in each dataset are presented in Table 3 and Figure S6. The 3-year time-dependent AUCs of the total score in the internal and external validation sets were 0.863 (95% CI, 0.747–0.939) and 0.774 (95% CI, 0.678–0.853) for PFS and 0.830 (95% CI, 0.708–0.916) and 0.776 (95% CI, 0.679–0.854) for overall survival (OS). In the internal and external validation sets, the AUC of the total score was statistically significantly higher than the IPI, KPI, and PINK (all $p < 0.05$) in predicting both the 3-year PFS status and 3-year OS status.

In the external validation set, the C-index of the total score for predicting PFS was significantly higher than IPI, KPI, and PINK (C-index [95% CI]: total score, 0.774 [0.711–0.838] vs. IPI, 0.617 [0.483–0.752], $p = 0.010$; vs. KPI, 0.678 [0.568–0.789], $p = 0.027$; vs. PINK, 0.649 (0.538–0.759), $p = 0.014$). For predicting OS, the total score outperformed IPI ($p = 0.045$) and KPI ($p = 0.005$) but failed to present a statistical difference in comparison to PINK ($p = 0.097$). The C-indexes are shown in Table S2.

Furthermore, the calibration curves of the total score for 3-year PFS prediction demonstrated a promising agreement between the actual survival probability and the predicted ones (Figure S7).

Early-stage patients in high-risk group according to MRI score benefited from early radiotherapy

Analysis was performed to test if the AI-based prognostic system could assist with deciding on the use of early radiotherapy ($\leq$3 cycles of chemotherapy before radiotherapy) or late radiotherapy (>3 cycles of chemotherapy before radiotherapy) for patients with early-stage NKTCL who received L-asparaginase/pegaspargase-based chemo-radiotherapy in the SYSUCC dataset ($n = 122$).

By using the X-tile software, the cutoff value of the MRI score that categorized patients in the training dataset into low risk ($n = 68$) vs. high risk ($n = 66$) was 57.14 (Figure S8A). No statistically significant difference was observed in the baseline characteristics between patients who underwent early and late radiotherapy in both the low- and high-risk groups (Table S3). For high-risk patients ($n = 48$), PFS was longer in the patients receiving early radiotherapy than those receiving late radiotherapy (hazard ratio [HR] = 0.414; 95% CI, 0.190–0.900; $p = 0.022$), while no difference was observed in low-risk patients ($n = 74$; HR = 0.902; 95% CI, 0.285–2.857; $p = 0.861$) (Figures 3A and 3B). Interestingly, classification of the patients into high risk vs. low risk based on the total score (Figure S8B) failed to discriminate between patients receiving early vs. late radiotherapy (Figures 3C and 3D).

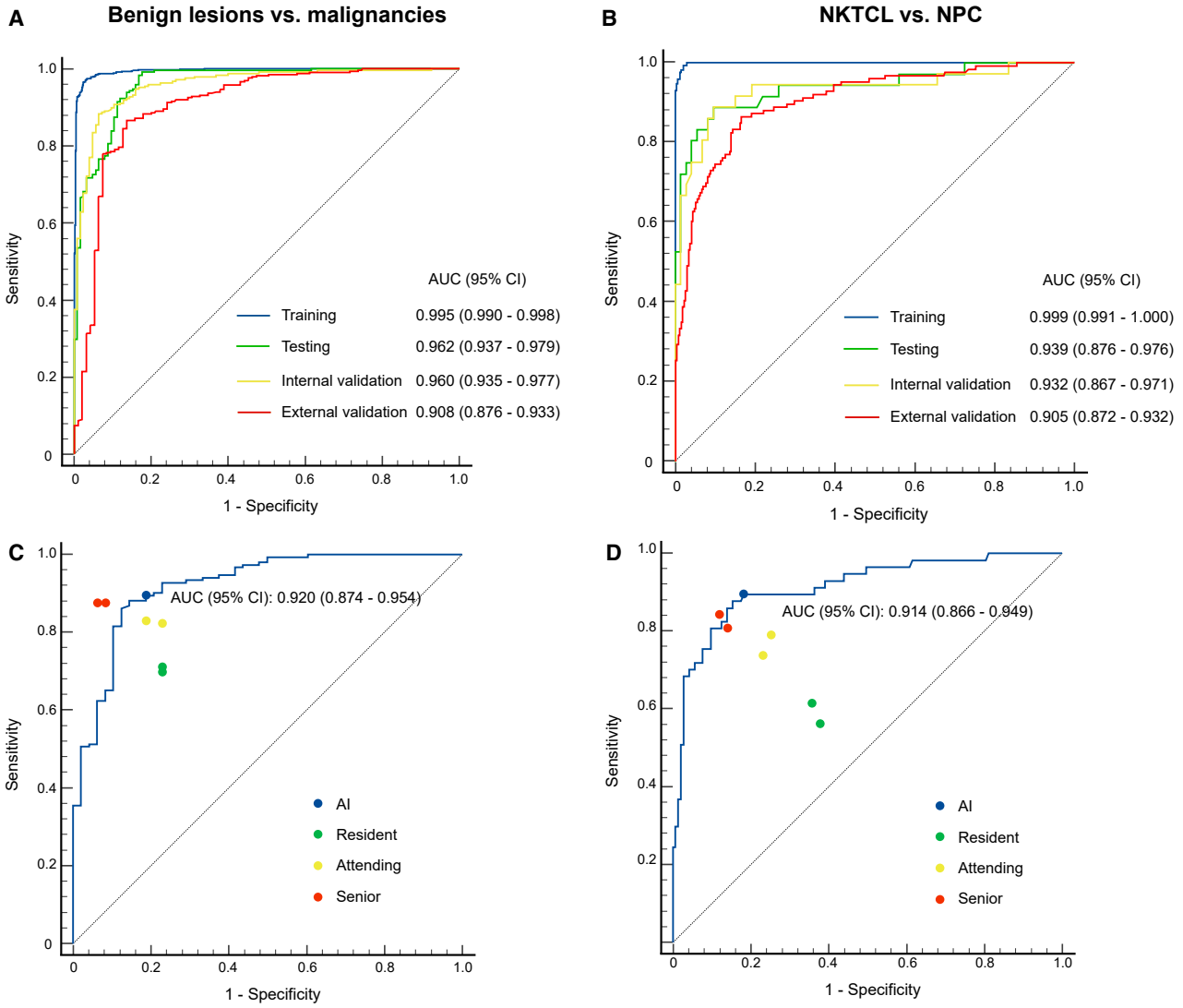


Figure 2. Performance of the diagnostic system

(A and B) The ROC curves of the AI systems in (A) detecting malignant lesions and (B) differential diagnosis of NKTCL and NPC in different datasets.

(C and D) Comparison between the AI systems and radiologists with different seniorities in (C) detecting malignant lesions and (D) differential diagnosis of NKTCL and NPC in two subsets, each with 200 cases randomly selected from the internal and external validation sets. AI, artificial intelligence; AUC, area under curve; NKTCL, natural killer/T cell lymphoma; NPC, nasopharyngeal carcinoma; ROC, receiver operating characteristic.

DISCUSSION

In this nationwide multicenter study, we constructed and validated an AI-based diagnostic system for the detection of nasopharyngeal malignancies and discriminating NKTCL from NPC using nasopharynx MRI images from a cohort of 3,088 patients. The diagnostic system showed satisfactory performance with significantly higher accuracy than a resident radiologist and similar accuracy to a senior radiologist. Moreover, the prognostic system for NKTCL was developed to improve the precision of survival prediction, which also had potential to assist with deciding treatment strategies for early-stage patients.

The onset of nasopharyngeal lesions is insidious, and the early symptoms are usually unnoticeable. There is a need to develop an efficient and valid tool to evaluate nasopharyngeal lesions for a preliminary differential diagnosis. Recently, several studies integrated convolutional neural networks and MRI to discriminate between nasopharyngeal malignant and benign lesions.[19–21] However, other subtypes of malignancies were not included other than NPC, restricting the clinical application of these techniques. The AI system built in the current study showed significantly better diagnostic accuracy than a resident radiologist and comparable accuracy to a senior radiologist and thus may be used to assist less experienced radiologists in differential diagnosis of nasopharyngeal lesions. Based on our findings, non-invasive nasopharynx MRI can provide complementary information to the existing diagnostic procedure. In the future, results of AI analysis may become a part of routine

Table 3. Three-year time-dependent AUCs for predicting PFS and OS

	PFS, AUC (95% CI)	*p* value	OS, AUC (95% CI)	*p* value
Training set (*n* = 134)				
Total score	0.875 (0.807–0.926)	referent	0.832 (0.758–0.891)	referent
MRI score	0.854 (0.783–0.909)	0.490	0.785 (0.706–0.851)	0.227
IPI	0.568 (0.479–0.653)	<0.001	0.582 (0.494–0.667)	<0.001
KPI	0.660 (0.573–0.739)	<0.001	0.655 (0.568–0.735)	<0.001
PINK	0.639 (0.552–0.720)	<0.001	0.640 (0.553–0.721)	<0.001
Internal validation set (*n* = 58)				
Total score	0.863 (0.747–0.939)	referent	0.830 (0.708–0.916)	referent
MRI score	0.812 (0.688–0.903)	0.334	0.771 (0.642–0.871)	0.307
IPI	0.573 (0.436–0.702)	<0.001	0.565 (0.428–0.695)	<0.001
KPI	0.622 (0.485–0.746)	<0.001	0.634 (0.497–0.757)	<0.001
PINK	0.620 (0.483–0.744)	<0.001	0.597 (0.460–0.723)	<0.001
External validation set (*n* = 96)				
Total score	0.774 (0.678–0.853)	referent	0.776 (0.679–0.854)	referent
MRI score	0.718 (0.617–0.805)	0.111	0.749 (0.651–0.832)	0.515
IPI	0.560 (0.455–0.661)	<0.001	0.609 (0.504–0.707)	0.006
KPI	0.645 (0.541–0.740)	0.019	0.578 (0.473–0.678)	<0.001
PINK	0.597 (0.492–0.696)	0.003	0.646 (0.542–0.741)	0.029

AUC, area under the curve; CI, confidence interval; IPI, the international prognostic index; KPI, the Korean prognostic index; MRI, magnetic resonance imaging; OS, overall survival; PINK, the prognostic index of natural killer lymphoma; PFS, progression-free survival.

radiological reports and potentially assist with decision-making by clinicians. Once the AI model is further optimized in expanded patient cohorts and applied in clinical practice in the near future, it has potential value in reducing misdiagnoses and missed diagnoses and thus in saving patients' time and monetary costs.

Several previous studies attempted to use radiomics to differentiate nasopharyngeal lymphoma from NPC.[22–24] One study used diffusion-weighted imaging and achieved an AUC of 0.768.[22] The other two studies used dynamic contrast-enhanced MRI and diffusion-weighted imaging and achieved AUCs of 0.866 and 0.961, respectively.[23,24] However, these studies had relatively small sample sizes and lacked independent validation. The AI-based diagnostic system constructed in the current study achieved AUCs of 0.932 and 0.905 in the independent internal and external validation sets, respectively. There are several advantages to the AI system. First, the apparent AUC is not much higher than reported previously but was based on independent validation using data from multiple centers, indicating better generalizability. Second, the model was based on automatically extracted MRI images without additional manual segmentation work, possessing higher applicability in daily practice. Third, our system only uses regular MRI scanning sequences (T1 weighted contrast enhanced), without using other sequences such as diffusion-weighted imaging.

NKTCL is characterized by extranodal involvement, and most patients are in the early stage when initially diagnosed. Accordingly, the predictive power in predicting patient prognosis with the conventional Ann Arbor staging system and the IPI is limited.[7,25] New prognostic models such as the KPI and PINK have been constructed specifically for NKTCL; however, there is no objective measure for primary nasal lesions.[26,27] In the current study, the AI prognostic system outperformed the IPI, KPI, and PINK in 3-year time-dependent AUC for predicting both PFS and OS and could serve as a complement to current clinical prognostic and risk stratification methods. Several reasons might contribute to the improved performance. First, nasopharynx MRI could estimate the invasion of the primary tumor, which is an important factor that affects survival outcomes and treatment response. Second, the application of radiomics technique can maximize the use of MRI information, contributing to the improvement of predictive efficiency. Moreover, the machine learning algorithm (random survival forest model) can process non-linear relationships and thus may be more suitable for processing complex associations than Cox proportional hazard regression.

Combined modality therapy, i.e., combination treatment of chemotherapy and radiotherapy, has been increasingly emphasized in the treatment of early-stage NKTCL.[28] However, the optimal therapeutic strategy, particularly the timing of radiotherapy, has not been established. Several previous studies attempted to identify the patient subpopulation that could benefit from early radiotherapy. For example, Qi and colleagues developed a nomogram-revised risk index and showed that early radiotherapy prolonged PFS in the subpopulation of non-complete response patients with intermediate/high risk but not in patients with low risk.[29] Another study by Zang and colleagues reported that early radiotherapy improved survival outcomes in patients with stage I–II nasal-type NKTCL.[30] The AI model developed in the current study expanded the pool of tools to predict the therapeutic response to early radiotherapy. Interestingly, in the current study, we found an association between longer PFS with early radiotherapy only in high-risk patients. We speculate that a higher MRI score, as derived from MRI images, may reflect a larger or deeper invasion of the primary lesions. We

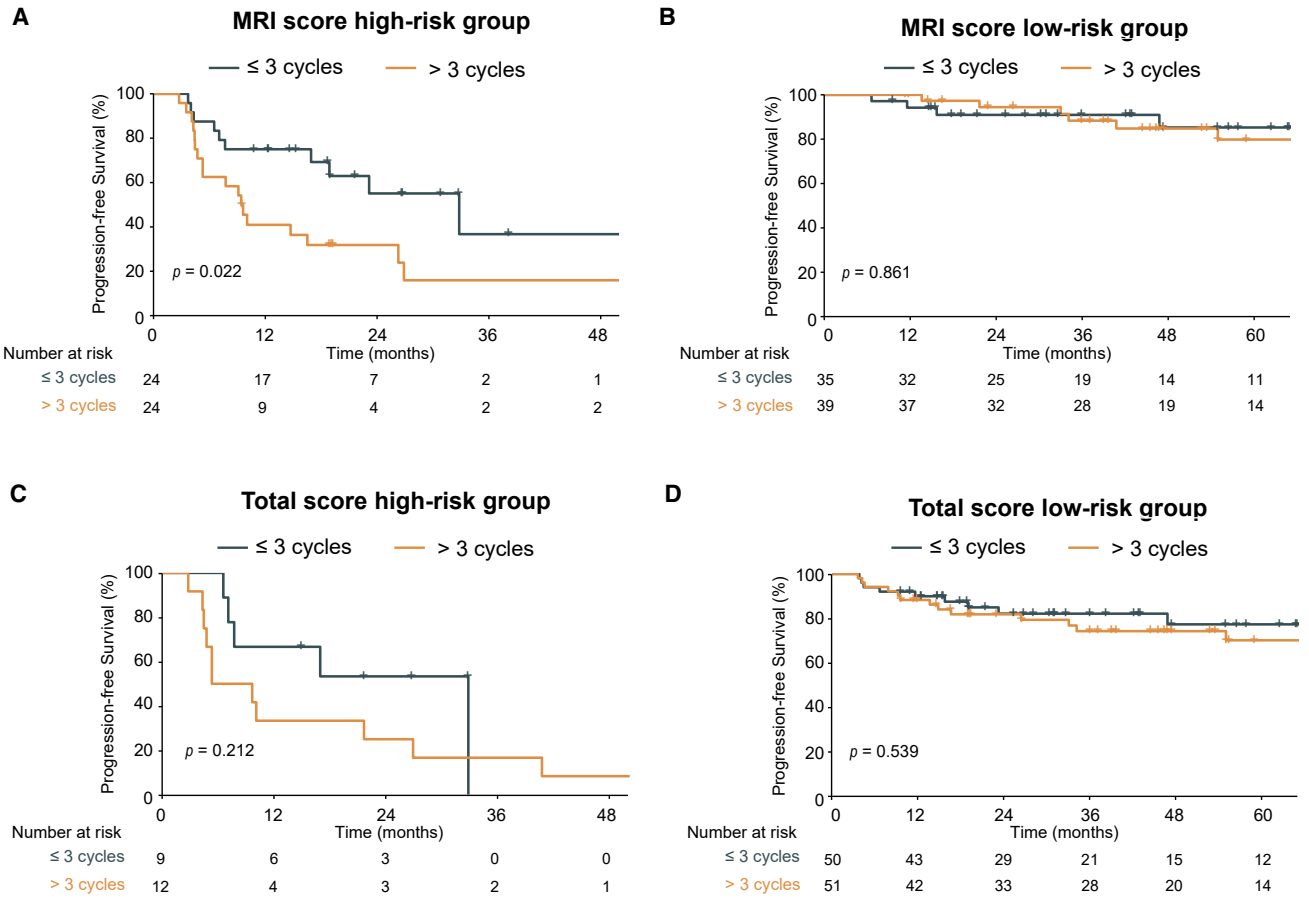


Figure 3. Kaplan-Meier estimate of PFS in patients with early-stage NKTCL receiving early or late radiotherapy

(A and B) PFS curves of (A) high-risk patients and (B) low-risk patients based on stratification of MRI score.

(C and D) PFS curves of (C) high-risk patients and (D) low-risk patients based on stratification of total score. Early radiotherapy is defined as receiving ≤3 cycles of chemotherapy before radiotherapy. Late radiotherapy is defined as receiving >3 cycles of chemotherapy before radiotherapy. MRI, magnetic resonance imaging; NKTCL, natural killer/T cell lymphoma; PFS, progression-free survival. *p* values were calculated by log rank test.

speculate that the model constructed in the current study could be useful in the decision-making of combined chemo-radiotherapy in patients with stage I–II NKTCL. For example, after a patient's nasopharyngeal MRI images are delineated, the model can provide the risk score of this case. High-risk patients, as classified based on this model, are more likely to benefit from early radiotherapy; in other words, indicating a short-course chemotherapy might be recommended. Future studies with a prospective design are needed to test such a speculation.

Based on the accuracy of the AI diagnostic and prognostic systems, we are now generating an online analysis system—NAIS (natural killer/T cell lymphoma artificial intelligence system; http://naids.sysucc.org.cn/). This user-friendly website assists clinicians with making radiological diagnoses of nasopharyngeal lesions and provides prognosis prediction for patients with NKTCL.

Limitations of the study

The current study has several limitations. First, the modeling was based on retrospective collection and thus is prone to substantial bias. Second, the prognostic system needs manual labeling of the nasopharyngeal lesions, which in turn represents a source of bias as well as a barrier for use in daily practice. Third, all patients were from a Chinese cohort, and further investigation is needed to validate its use in other populations.

Conclusion

In this work, the diagnostic system based on AI algorithms and MRI images may provide assistance in the differential diagnosis of nasopharyngeal lesions. The prognostic system could improve the precision of prognosis prediction and have the potential to give implications for therapeutic preferences for patients with early-stage NKTCL.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCES AVAILABILITY
  - Lead contact

- Materials availability
- Data and code availability
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Participants and study design
- METHOD DETAILS
  - MRI scanning protocol
  - The preprocessing of MRI images
  - Development of the diagnostic system
  - Development of the prognostic system
  - Model validation
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xcrm.2024.101551.

ACKNOWLEDGMENTS

This work was supported by grants from the National Key Research and Development Program (2022YFC2502602), the National Natural Science Foundation of China (82230001 and 82270199), the Guangzhou Science and Technology Program (2024B03J1291), the Sun Yat-Sen University Clinical Research 5010 Program (2020009), and the Beijing Xisike Clinical Oncology Research Foundation (Y-XD2019-124 and Y-SY2021ZD-0110). The authors thank Kehong Zhang, MD, PhD, from Ivy Medical Editing (Shanghai, China) for revising the manuscript. The graphical abstract was created with BioRender.com.

AUTHOR CONTRIBUTIONS

Conceptualization, Q.C., C.L., Y.Z., and Y. Deng; formal analysis, methodology, and visualization, Y.Z., Y. Deng, Q.Z., and B.J.; writing – original draft, Y.Z. and Y. Deng; data curation and writing – review & editing, all authors; supervision, Q.C., C.L., L.W., and C.X.; funding acquisition, Q.C.

DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: November 23, 2023
Revised: March 5, 2024
Accepted: April 11, 2024
Published: May 1, 2024

REFERENCES

1. Kimura, H., de Leval, L., Cai, Q., and Kim, W.S. (2022). EBV-associated NK and T-cell lymphoid neoplasms. Curr. Opin. Oncol. *34*, 422–431. https://doi.org/10.1097/cco.0000000000000889.
2. Kim, T.M., Lee, S.Y., Jeon, Y.K., Ryoo, B.Y., Cho, G.J., Hong, Y.S., Kim, H.J., Kim, S.Y., Kim, C.S., Kim, S., et al. (2008). Clinical heterogeneity of extranodal NK/T-cell lymphoma, nasal type: a national survey of the Korean Cancer Study Group. Ann. Oncol. *19*, 1477–1484. https://doi.org/10.1093/annonc/mdn147.
3. El-Naggar, A., Chan, J., Grandis, J., Takata, T., and Slootweg, P. (2017). WHO Classification of Head and Neck Tumours. WHO Classification of Tumours *9*. 4th Edition.
4. Cai, Q., Cai, J., Fang, Y., and Young, K.H. (2019). Epstein-Barr Virus-Positive Natural Killer/T-Cell Lymphoma. Front. Oncol. *9*, 386. https://doi.org/10.3389/fonc.2019.00386.
5. Chen, Y.P., Chan, A.T.C., Le, Q.T., Blanchard, P., Sun, Y., and Ma, J. (2019). Nasopharyngeal carcinoma. Lancet (London, England) *394*, 64–80. https://doi.org/10.1016/s0140-6736(19)30956-0.
6. Yamaguchi, M., Suzuki, R., and Oguchi, M. (2018). Advances in the treatment of extranodal NK/T-cell lymphoma, nasal type. Blood *131*, 2528–2540. https://doi.org/10.1182/blood-2017-12-791418.
7. Tse, E., and Kwong, Y.L. (2017). The diagnosis and management of NK/T-cell lymphomas. J. Hematol. Oncol. *10*, 85. https://doi.org/10.1186/s13045-017-0452-9.
8. King, A.D., Lei, K.I.K., Richards, P.S., and Ahuja, A.T. (2003). Non-Hodgkin's lymphoma of the nasopharynx: CT and MR imaging. Clin. Radiol. *58*, 621–625. https://doi.org/10.1016/s0009-9260(03)00182-x.
9. King, A.D., Wong, L.Y.S., Law, B.K.H., Bhatia, K.S., Woo, J.K.S., Ai, Q.Y., Tan, T.Y., Goh, J., Chuah, K.L., Mo, F.K.F., et al. (2018). MR Imaging Criteria for the Detection of Nasopharyngeal Carcinoma: Discrimination of Early-Stage Primary Tumors from Benign Hyperplasia. AJNR. Am. J. Neuroradiol. *39*, 515–523. https://doi.org/10.3174/ajnr.A5493.
10. Trister, A.D., Buist, D.S.M., and Lee, C.I. (2017). Will Machine Learning Tip the Balance in Breast Cancer Screening? JAMA Oncol. *3*, 1463–1464. https://doi.org/10.1001/jamaoncol.2017.0473.
11. Yu, K.H., Beam, A.L., and Kohane, I.S. (2018). Artificial intelligence in healthcare. Nat. Biomed. Eng. *2*, 719–731. https://doi.org/10.1038/s41551-018-0305-z.
12. Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L.H., and Aerts, H.J.W.L. (2018). Artificial intelligence in radiology. Nat. Rev. Cancer *18*, 500–510. https://doi.org/10.1038/s41568-018-0016-5.
13. Huynh, E., Hosny, A., Guthier, C., Bitterman, D.S., Petit, S.F., Haas-Kogan, D.A., Kann, B., Aerts, H.J.W.L., and Mak, R.H. (2020). Artificial intelligence in radiation oncology. Nat. Rev. Clin. Oncol. *17*, 771–781. https://doi.org/10.1038/s41571-020-0417-8.
14. Campanella, G., Hanna, M.G., Geneslaw, L., Miraflor, A., Werneck Krauss Silva, V., Busam, K.J., Brogi, E., Reuter, V.E., Klimstra, D.S., and Fuchs, T.J. (2019). Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. Nat. Med. *25*, 1301–1309. https://doi.org/10.1038/s41591-019-0508-1.
15. Luo, H., Xu, G., Li, C., He, L., Luo, L., Wang, Z., Jing, B., Deng, Y., Jin, Y., Li, Y., et al. (2019). Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol. *20*, 1645–1654. https://doi.org/10.1016/s1470-2045(19)30637-0.
16. Guo, R., Hu, X., Song, H., Xu, P., Xu, H., Rominger, A., Lin, X., Menze, B., Li, B., and Shi, K. (2021). Weakly supervised deep learning for determining the prognostic value of (18)F-FDG PET/CT in extranodal natural killer/T cell lymphoma, nasal type. Eur. J. Nucl. Med. Mol. Imag. *48*, 3151–3161. https://doi.org/10.1007/s00259-021-05232-3.
17. Girum, K.B., Rebaud, L., Cottereau, A.S., Meignan, M., Clerc, J., Vercellino, L., Casasnovas, O., Morschhauser, F., Thieblemont, C., and Buvat, I. (2022). (18)F-FDG PET maximum intensity projections and artificial intelligence: a win-win combination to easily measure prognostic biomarkers in DLBCL patients. J. Nucl. Med. *63*, 1925–1932. https://doi.org/10.2967/jnumed.121.263501.
18. Kong, H., Zhu, H., Zheng, X., Jiang, M., Chen, L., Lan, L., Ren, J., Luo, X., Zheng, J., Zheng, Z., et al. (2022). Machine Learning Models for the Diagnosis and Prognosis Prediction of High-Grade B-Cell Lymphoma. Front. Immunol. *13*, 919012. https://doi.org/10.3389/fimmu.2022.919012.
19. Ke, L., Deng, Y., Xia, W., Qiang, M., Chen, X., Liu, K., Jing, B., He, C., Xie, C., Guo, X., et al. (2020). Development of a self-constrained 3D DenseNet model in automatic detection and segmentation of nasopharyngeal carcinoma using magnetic resonance images. Oral Oncol. *110*, 104862. https://doi.org/10.1016/j.oraloncology.2020.104862.
20. Wong, L.M., King, A.D., Ai, Q.Y.H., Lam, W.K.J., Poon, D.M.C., Ma, B.B.Y., Chan, K.C.A., and Mo, F.K.F. (2021). Convolutional neural network for discriminating nasopharyngeal carcinoma and benign hyperplasia on MRI. Eur. Radiol. *31*, 3856–3863. https://doi.org/10.1007/s00330-020-07451-y.

21. Deng, Y., Li, C., Lv, X., Xia, W., Shen, L., Jing, B., Li, B., Guo, X., Sun, Y., Xie, C., and Ke, L. (2022). The contrast-enhanced MRI can be substituted by unenhanced MRI in identifying and automatically segmenting primary nasopharyngeal carcinoma with the aid of deep learning models: An exploratory study in large-scale population of endemic area. Comput. Methods Progr. Biomed. *217*, 106702. https://doi.org/10.1016/j.cmpb.2022.106702.

22. Lian, S., Zhang, C., Chi, J., Huang, Y., Shi, F., and Xie, C. (2020). Differentiation between nasopharyngeal carcinoma and lymphoma at the primary site using whole-tumor histogram analysis of apparent diffusion coefficient maps. Radiol. Med. *125*, 647–653. https://doi.org/10.1007/s11547-020-01152-8.

23. Yu, X.P., Hou, J., Li, F.P., Xiang, W., Lu, Q., Hu, Y., and Wang, H. (2016). Quantitative dynamic contrast-enhanced and diffusion-weighted MRI for differentiation between nasopharyngeal carcinoma and lymphoma at the primary site. Dentomaxillofacial Radiol. *45*, 20150317. https://doi.org/10.1259/dmfr.20150317.

24. Song, C., Cheng, P., Cheng, J., Zhang, Y., Sun, M., Xie, S., and Zhang, X. (2020). Differential diagnosis of nasopharyngeal carcinoma and nasopharyngeal lymphoma based on DCE-MRI and RESOLVE-DWI. Eur. Radiol. *30*, 110–118. https://doi.org/10.1007/s00330-019-06343-0.

25. Cheson, B.D., Fisher, R.I., Barrington, S.F., Cavalli, F., Schwartz, L.H., Zucca, E., and Lister, T.A.Alliance, Australasian Leukaemia and Lymphoma Group; Eastern Cooperative Oncology Group; European Mantle Cell Lymphoma Consortium (2014). Recommendations for initial evaluation, staging, and response assessment of Hodgkin and non-Hodgkin lymphoma: the Lugano classification. J. Clin. Oncol. *32*, 3059–3068. https://doi.org/10.1200/jco.2013.54.8800.

26. Lee, J., Suh, C., Park, Y.H., Ko, Y.H., Bang, S.M., Lee, J.H., Lee, D.H., Huh, J., Oh, S.Y., Kwon, H.C., et al. (2006). Extranodal natural killer T-cell lymphoma, nasal-type: a prognostic model from a retrospective multicenter study. J. Clin. Oncol. *24*, 612–618. https://doi.org/10.1200/jco.2005.04.1384.

27. Kim, S.J., Yoon, D.H., Jaccard, A., Chng, W.J., Lim, S.T., Hong, H., Park, Y., Chang, K.M., Maeda, Y., Ishida, F., et al. (2016). A prognostic index for natural killer cell lymphoma after non-anthracycline-based treatment: a multicentre, retrospective analysis. Lancet Oncol. *17*, 389–400. https://doi.org/10.1016/s1470-2045(15)00533-1.

28. Kim, S.J., Yoon, S.E., and Kim, W.S. (2018). Treatment of localized extranodal NK/T cell lymphoma, nasal type: a systematic review. J. Hematol. Oncol. *11*, 140. https://doi.org/10.1186/s13045-018-0687-0.

29. Qi, S.N., Yang, Y., Zhang, Y.J., Huang, H.Q., Wang, Y., He, X., Zhang, L.L., Wu, G., Qu, B.L., Qian, L.T., et al. (2020). Risk-based, response-adapted therapy for early-stage extranodal nasal-type NK/T-cell lymphoma in the modern chemotherapy era: A China Lymphoma Collaborative Group study. Am. J. Hematol. *95*, 1047–1056. https://doi.org/10.1002/ajh.25878.

30. Zang, J., Li, C., Luo, S.Q., Wang, J.H., Xu, M., Zhao, L.N., Li, W.W., Yang, H., Xiao, F., Hitchcock, Y.J., and Shi, M. (2015). Early radiotherapy has an essential role for improving survival in patients with stage I-II nasal-type of NK/T cell lymphoma treated with L-asparaginase-containing chemotherapy–a single institution experience. Ann. Hematol. *94*, 583–591. https://doi.org/10.1007/s00277-014-2244-4.

31. Swerdlow, S., Campo, E., Harris, N., Jaffe, E., Pileri, S., Stein, H., and Thiele. (2017). J.WHO Classification of Tumours of Haematopoietic and Lymphoid Tissues. WHO Classification of Tumours *2*.

32. Zhang, Y., Ma, S., Cai, J., Yang, Y., Jing, H., Shuang, Y., Peng, Z., Li, B., Liu, P., Xia, Z., et al. (2021). Sequential P-GEMOX and radiotherapy for early-stage extranodal natural killer/T-cell lymphoma: A multicenter study. Am. J. Hematol. *96*, 1481–1490. https://doi.org/10.1002/ajh.26335.

33. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. (2021). Asymmetric Loss For Multi-Label Classification. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 82–91. https://doi.org/10.1109/ICCV48922.2021.00015.

34. Perez-Garcia, F., Sparks, R., and Ourselin, S. (2021). TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Article. Comput. Methods Progr. Biomed. 208106236. https://doi.org/10.1016/j.cmpb.2021.106236.

35. Fortin, J.-P., Cullen, N., Sheline, Y.I., Taylor, W.D., Aselcioglu, I., Cook, P.A., Adams, P., Cooper, C., Fava, M., McGrath, P.J., et al. (2018). Harmonization of cortical thickness measurements across scanners and sites. Article. *Neuroimage*. *167*, 104–120. https://doi.org/10.1016/j.neuroimage.2017.11.024.

36. Kursa, M.B., and Rudnicki, W.R. (2010). Feature Selection with the Boruta Package. Article. J. Stat. Software *36*, 1–13. https://doi.org/10.18637/jss.v036.i11.

37. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., and Lauer, M.S. (2008). RANDOM SURVIVAL FORESTS. Article. Ann. Appl. Stat. *2*, 841–860. https://doi.org/10.1214/08-aoas169.

38. (1993). A predictive model for aggressive non-Hodgkin's lymphoma. N. Engl. J. Med. *329*, 987–994. https://doi.org/10.1056/nejm199309303291402.

39. Cai, Q., Luo, X., Liang, Y., Rao, H., Fang, X., Jiang, W., Lin, T., Lin, T., and Huang, H. (2013). Fasting blood glucose is a novel prognostic indicator for extranodal natural killer/T-cell lymphoma, nasal type. Br. J. Cancer *108*, 380–386. https://doi.org/10.1038/bjc.2012.566.

40. Heagerty, P.J., Lumley, T., and Pepe, M.S. (2000). Time-dependent ROC curves for censored survival data and a diagnostic marker. Biometrics *56*, 337–344. https://doi.org/10.1111/j.0006-341x.2000.00337.x.

41. Harrell, F.E., Jr., Califf, R.M., Pryor, D.B., Lee, K.L., and Rosati, R.A. (1982). Evaluating the yield of medical tests. JAMA *247*, 2543–2546.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Python 3.6.10	Python	https://www.python.org
Pytorch 1.4.0+cu10.1	Pytorch	https://pytorch.org
Torchvision 0.5.0+cu10.1	Pytorch	https://pytorch.org
Numpy 1.19.2	Numpy	https://numpy.org
TorchIO 0.18.75	TorchIO	https://torchio.readthedocs.io/
Opencv-python 4.2.0.34	Opencv	https://docs.opencv.org
Pillow 8.3.1	Pillow	https://pillow.readthedocs.io
R 3.6.1	The R Project for Statistical Computing	https://www.r-project.org
SPSS software 25.0	IBM	https://www.ibm.com/spss
MedCalc software 20.0.22	MedCalc software Ltd	https://www.medcalc.org/
X-tile software	Yale School Of Medicine	https://medicine.yale.edu/lab/rimm/research/software/
BioRender	BioRender website	https://www.biorender.com/

RESOURCES AVAILABILITY

Lead contact

Further information and resource requests should be directly to and will be fulfilled by the lead contract, QingQing Cai (caiqq@sysucc.org.cn).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- To protect patient privacy, the MRI images used in this study are not publicly accessible. Deidentified patient-level clinical and survival data reported in this paper will be shared by the lead contact upon reasonable request.
- This paper does not report the original code.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Participants and study design

This multicenter, retrospective cohort study was conducted in 9 hospitals from China, included Sun Yat-sen University Cancer Center (SYSUCC), Beijing TongRen Hospital, Peking University First Hospital, the First People's Hospital of Foshan, Fujian Provincial Cancer Hospital & Institute, Guangdong Provincial People's Hospital, the Second Affiliated Hospital of Suzhou University, Sun Yat-sen Memorial Hospital, and the First Affiliated Hospital of Sun Yat-sen University. Data of patients with abnormality of nasopharynx in contrast-enhanced MRI images prior to treatment were retrospectively collected. MRI was conducted using either 1.5T or 3.0T MR machines, and were stored in Digital Imaging and Communications in Medicine (DICOM) format. For patients with malignant lesions, MRI scanning must be conducted within one month prior to anti-tumor treatment for inclusion in the final analysis. All included patients had undergone biopsy of nasopharyngeal lesions. Malignant lesions was categorized according to the World Health Organization classification (the revised 4th edition of hematopoietic and lymphoid tissues,[31] and the 4th edition of head and neck tumors[3]) based on pathological report at the participating hospitals. Extensive primary tumor was defined as the primary lesion that extending to adjacent structures.[32]

A total of 3276 patients who met the eligibility criteria were screened. Patients without definitive pathological diagnosis (n = 16), underwent biopsy of nasopharyngeal lesions before nasopharynx MRI (n = 98) or with a history of other malignancies (n = 23) were excluded. A quality control procedure was performed with the MRI images: images not containing the entire lesion (n = 5), with insufficient resolution (n = 40; e.g., motion artifacts) or inconsistent number of layers (n = 6) were excluded. Finally, 3088 cases were used for further development of diagnostic system.

For the task of discrimination of benign and malignant lesions, 2507 patients from SYSUCC during a period from Jan, 2019 to Dec, 2021 were used; these patients were randomly allocated, at a ratio of 7: 1.5: 1.5, to a training (n = 1761), a testing (n = 373), and an internal validation set (n = 373). For external validation, 430 patients from 8 other hospitals were included. For discrimination of NKTCL and NPC, an expanded cohort of NKTCL patients (71 from SYSUCC and 80 from other hospitals) between Jan, 2013 to Dec, 2018 was included. A total of 244 NKTCL patients and 490 NPC patients (randomly selected at a ratio of 1:2) from SYSUCC were allocated at a ratio of 7: 1.5: 1.5, to a training (n = 516), a testing (n = 109), and an internal validation set (n = 109). Patients from 8 other hospitals (n = 398) were used for external validation.

The development of prognostic system was conducted in 288 NKTCL patients. Patients who had incomplete baseline characteristic data or had <6 months follow-up with no PFS event occurred were excluded. The patients from SYSUCC (n = 192) were randomly allocated at a ratio of 7: 3 to a training set (n = 134) and an internal validation set (n = 58). Patients from 8 other centers (n = 96) were used for external validation.

This study was approved by the ethics committee of each participating hospital, and was conducted in compliance with the Declaration of Helsinki. The requirement for written informed consent was waived because of the retrospective nature of the study and only pre-existing medical data were used.

METHOD DETAILS

MRI scanning protocol

All patients underwent pretreatment head or head & neck MR examination. The acquisition parameters of MRI scanning are presented in Table S4. The scanning range of head MRI was set from the suprasellar cistern to the level of epiglottis. The scanning range of head and neck MRI was set from the suprasellar cistern to the inferior border of the sternal end of clavicle. T1-and T2-weighted fast spin echo images were acquired in the axial plane prior to contrast material injection. T1-weighted enhanced sequence was conducted 40 s following intravenous administration of Gd-DTPA at a dosage of 0.1 mmol/kg

The preprocessing of MRI images

The input 3D image consists of 16 slices, covering the structures of the nasopharynx. To meet the input requirements of the model, all MR images were sampled using bilinear interpolation to a size of 384*384 pixels. Before being fed into the model, each images undergo a standardization process using the following steps.

(1) Calculate the pixel mean value "u" of the images

$$u = \frac{1}{N}\sum_{k=0}^{N} X_k$$

X represents the set of pixels in the image, N represents the total number of pixels in the image, and X_k represents the value of the ith pixel.

(2) Calculate the pixel mean value "m" of the foreground of the images

$$m = \frac{1}{\sum_{k}^{N} 1_{(X_k > u)}} \sum_{k=0}^{N} X_{k(X_k > u)}$$

u represents the average value of pixels in the image.

(3) Obtain the standardized images "Z"

$$Z = \frac{X}{m}$$

m represents the average value of pixels in the foreground of the image.

Development of the diagnostic system

The 3D DenseNet model was used to identify the features in non-segmented, axial, T1-weighted contrast-enhanced MRI images (Figure S9). The 3D DenseNet model used an end-to-end two-class tasking learning network architecture. In order to obtain rich and representative features, the 3D DenseNet has 6 Denseblock and 5 Transitionblock. Each Denseblock contained several densely

connected convolutional layers, and each convolutional layer (conv) was followed by a Batch Normalization layer (BN)and the Leaky Rectified Linear Unit activation (LeakyReLU). DenseNet employs 1 × 1×1 convolution as bottleneck layer in DenseBlock structure [BN + LeakyReLU + conv1×1 × 1 + BN + LeakyReLU + conv3×3 × 3]. Transitionblock consisted of a convolutional layer and a pooling layer. The Transition Block typically consists of multiple layers, including an activation layer, a Batch Normalization layer, a convolutional layer, and a pooling layer. The structure is defined as [BN + LeakyReLU + conv1×1 × 1 + maxpool]. The model takes the nasopharynx 3D MR image as input and provides the diagnostic probability as output.

Pytorch (version 1.4.0) was used to train the 3D DenseNet model with 2 GPUs. The parameter settings are shown in Table S5. Optimization of the total loss of asymmetric loss of the 3D DenseNet during training process was performed using the 'Adam' optimizer, with minibatch size of 6 and the initial learning rate of 1e−4.[33] The probability margin, positive and negative focusing parameters of asymmetric loss were 0.2, 1 and 2, respectively. The MRI images were augmented using RandomAffine (rotation, scaler, translation) during the training process.[34] An adaptive learning rate scheduling strategy was used for training. Separated training and testing sets were used to train and fine-tune model performance. The learning rate was reduced by half when the loss of the testing set stopped decreasing for 5 epochs, and the monitoring threshold was set to 1e−4. The model with the best area under curve (AUC) in the testing set was selected for internal and external validation.

Development of the prognostic system

The regions of interest (ROIs) were manually delineated in axial, T1-weighted contrast-enhanced MRI images with the ITK-SNAP software (version 3.8.0) by a radiologist with 5-year experience, and was supervised and refined by a senior radiologist with 25-year experience. The delineation was finalized when the two radiologists reached a consensus.

Pyradiomics (version 3.0) was used to extract radiomic features of the following 7 types: first-order, three-dimensional shape, gray-level co-occurrence matrices, gray-level run-length matrix, gray level size-zone matrix, neighboring gray tone difference matrix, and gray -level dependence matrix. A total of 1595 radiomics features were extracted for each patient. To ensure the stability and reproducibility of radiomics features in segmentation, 563 features with variance less than 1e−4 were excluded from the modeling. ComBat harmonization was performed to eliminate nonbiological variations across different MRI scanners and centers.[35] To minimize overfitting and inefficient computation, feature selection was conducted based on a binary classification task in the subset of patients with definite 2-year PFS status (yes versus no) using the Boruta method.[36] A total of 23 features that represent the entire feature set were selected (Table S6).

An "MRI score" was obtained after entering the selected 23 features into a Random Survival Forest (RSF) model for PFS.[37] In addition to MRI score, the following 10 variables were included in the prediction model: age (≤60 vs. > 60 years), gender (male vs. female), B symptoms (yes vs. no), Eastern Cooperative Oncology Group-performance status (0–1 vs. ≥ 2), Ann Arbor staging (stage I-II vs. stage III-IV), extranodal involvement (0–1 vs. ≥ 2), fasting blood glucose (≤6.10 vs. > 6.10 mmol/L), lactate dehydrogenase (≤245 vs. > 245 U/L), regional lymph node involvement (yes vs. no), distant lymph node involvement (yes vs. no).[26,27,38,39] A "total score" was obtained after entering the MRI score and the other 10 clinical variables into the RSF model for PFS.

Model validation

Model performance was validated in an internal validation set from SYSUCC and an external validation set from 8 other hospitals. Two subsets of 200 cases each were randomly selected from the validation sets to compare the performance of the diagnostic systems to radiologists. Six radiologists, 2 at each of the 3 seniority levels, were asked to conduct image evaluation independently. The 2 residents (one Bachelor of Medicine and one Master of Medicine) had 1 and 2 years of experience in radiological diagnosis, respectively. The 2 attending radiologists (Doctor of Medicine) had 5 years of experience in radiological diagnosis. The 2 senior radiologists (Doctor of Medicine) had 18 and 25 years of experience in radiological diagnosis, respectively. The radiologists were not involved in patient enrollment, MR images annotation, and were blinded to patients' clinical information.

The performance of the NKTCL prognostic system was compared to several commonly used prognostic models in clinical practice, including the international prognostic index (IPI), the Korean prognostic index (KPI), and the prognostic index of natural killer lymphoma (PINK).[26,27,38]

QUANTIFICATION AND STATISTICAL ANALYSIS

Descriptive statistics were used to present baseline characteristics. Categorical variables are shown as number and percentage in each category. Area under the curve (AUC) in receiver operating characteristic (ROC) analysis, accuracy, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) were used to evaluate the performance of the diagnostic system. The 95% confidence interval (CI) was calculated by Clopper-Pearson method. The differences in diagnostic accuracy, sensitivity, specificity, PPV, and NPV were assessed between the AI system and radiologists using the Chi-square test. The time-dependent AUC[40] and concordance index (C-index)[41] were used to evaluate the performance of the NKTCL prognostic system. Calibration curves were used to evaluate if the predicted outcomes approximate the actual outcomes. PFS was defined as the time from pathological diagnosis to disease progression or death. Overall survival (OS) was defined as the time from pathological diagnosis to death from any cause. PFS and OS were estimated by Kaplan-Meier method and compared by log rank test. The optimal cut-off value of the MRI score and total score was calculated using X-tile software (v3.6.1; Yale University). Statistical analyses were performed using the R software (version 3.6.1), MedCalc software (version 20.0.22), and SPSS software (version 25.0). Two-sided P of <0.05 was considered statistically significant.