

Supplementary Materials

List

Part I. Supplementary Methods

Methods 1. Inclusion and exclusion criteria

Methods 2. Therapeutic regimens

Methods 3. Treatment response evaluation

Methods 4. Follow-up

Methods 5. Clinical classification

Methods 6. Imaging scanning protocol

Methods 7. Scan labelling

Supplementary References

Part II. Supplementary Figures

Fig. 1 Kaplan-Meier curves for locoregional relapse-free survival at the earliest visible recurrent nasopharyngeal carcinoma lesion and when recurrent nasopharyngeal carcinoma was confirmed.

Fig. 2 Distributions of the clinical staging of recurrence at the earliest visible recurrent nasopharyngeal carcinoma lesion and when recurrent nasopharyngeal carcinoma was confirmed.

Fig. 3 Dataset statistics.

Fig. 4 Receiver operating characteristic curves of the RAIN model (sequential scan model) vs single scan model for discriminating follow-up H&N MR scans in the training and tuning cohorts.

Fig. 5 Receiver operating characteristic curves of the RAIN model (sequential scan model) vs single scan model for discriminating follow-up H&N MR scans in the internal and external validation cohorts.

Fig. 6 Receiver operating characteristic curves of the RAIN model (sequential scan model) in discriminating follow-up H&N MR scans in the internal validation cohort stratified by different clinical factors.

Fig. 7 Receiver operating characteristic curves of the RAIN model (sequential scan model) in discriminating follow-up H&N MR scans in the external validation cohort stratified by different clinical factors.

Fig. 8 A successful case of a true positive in a 51-year-old male who presented with recurrence in the left fossa of rosenmüller.

Fig. 9 A successful case of a true positive in a 33-year-old male who presented with recurrence in the right parietal wall and right lateral wall of the nasopharynx involving the right palatine fossa, medial pterygoid, musculus longus capitis, foramen lacerum and cavernous sinus.

Fig. 10 A successful case of a true positive in a 46-year-old male who presented with recurrence in the left retropharyngeal lymph node.

Fig. 11 A failed case of a false positive in a 33-year-old male who presented with recurrence in his left paranasopharyngeal space involving the left retropharyngeal space, palatine fossa, foramen lacerum, oval foramen and cavernous sinus.

Fig. 12 A failed case of a false positive in a 59-year-old male presented with recurrence in his left cervical lymph node (IIb region).

Fig. 13 A failed case of a false negative in a 38-year-old female presented with recurrence in her right parietal wall and right lateral wall of the nasopharynx involving the right musculus levator veli palatini, musculus tensor veli palatini, musculus longus capitis, medial pterygoid, lateral pterygoid muscle, foramen lacerum, oval foramen and cavernous sinus.

Fig. 14 A failed case of a false negative in a 46-year-old male who presented with recurrence in the left cervical lymph node (IIb region).

Fig. 15 Scan review process.

Fig. 16 Flowchart of the RAIN model.

Fig. 17 Flowchart of the single scan model.

Part III. Supplementary Tables

Table 1. Imaging characteristics of the confirmed recurrences.

Table 2. Imaging characteristics of the earliest visible recurrences.

Table 3. Performance of the ResNet18 neural network for identifying scans with recurrence in the training and tuning cohorts.

Table 4. Performance and diagnostic accuracy of the RAIN model (sequential scan model) vs single scan model in identifying scans with confirmed recurrence in the retrospective study.

Table 5. Diagnostic accuracy of RAIN Model (sequential scan model) in identifying MR scans with recurrence in retrospective validation cohorts stratified by different clinical factors.

Table 6. Improvement of RAIN model (sequential scan model) on timely diagnosis of recurrence at individual level in the retrospective study.

Table 7. False positive alerts.

Table 8. Improvement of RAIN model (sequential scan model) and readers on timely diagnosis of recurrence at individual level in the prospective study.

Table 9. The acquisition parameters of H&N MR scanning.

Table 10. Labelling tool prompts and the instructions.

Table 11. Architecture of the model.

Part IV. Study Protocol

Part I. Supplementary Methods

Methods 1. Inclusion and exclusion criteria

For the cohorts employed in the construction and verification of RAIN, we retrieved patients from the Hospital 1 and Hospital 2.

Patients who met the following inclusion criteria were enrolled in this study: (i) aged ≥ 18 years with pathologically confirmed nasopharyngeal carcinoma (NPC) at the initial diagnosis; (ii) restaged to I-IVa ($T_{1-4}N_{0-3}M_0$) according to the American Joint Committee on Cancer (AJCC) 8th edition staging classification system ¹; (iii) treated with radical radiotherapy (RT) $\pm$ neoadjuvant or concurrent chemotherapy; (iv) imaging confirmation of complete remission (CR) of gross tumor at the primary site and metastatic lymph nodes in the cervical region within 6 months post RT ²; (v) pretreatment, posttreatment and regular follow-up (at least annual) head and neck (H&N) magnetic resonance (MR) examinations; and (vi) high-resolution H&N follow-up MR images, including axial T1-, T2-weighted imaging (T1-WI, T2-WI) and contrast-enhanced T1-weighted imaging (CE T1-WI). Patients who met any of the following exclusion criteria were excluded from this study: (i) distant metastasis at first diagnosis; (ii) did not receive radical RT or had an interruption to RT >7 days; (iii) absence T1-, T2-WI or CE T1-WI scans or had pretreatment or follow-up MR images of insufficient quality to obtain measurements (i.e., motion artefacts); (iv) unknown efficacy evaluation or partial remission (PR) or stable disease (SD) beyond 6 months after RT ²; (v) irregular follow-up surveys; or (vi) other malignancies or previous antitumor therapy.

Methods 2. Therapeutic regimens

All patients with non-disseminated NPC received radical RT to the primary tumor in the nasopharynx and metastatic lymph node(s) in the cervical region, if necessary. Radical RT was delivered as two-dimensional conventional RT (2D-CRT), three-dimensional conformal RT (3D-CRT) or intensity-modulated radiation therapy (IMRT). The requirement of the cumulative radiation dose was 66.0 Gy or greater in 30–35 fractions, with five daily fractions per week for 6–7 weeks. Platinum-based chemotherapy, including concurrent chemoradiotherapy (CCRT) every 3 weeks for 2–3 cycles or weekly for 6–7 cycles and neoadjuvant chemotherapy (NACT) every 3 weeks for 2–4 cycles, was implemented at the physician's discretion depending on the patient's physical status and the stage of disease.

Methods 3. Treatment response evaluation

Treatment response was evaluated on each patient's H&N MR images within 6 months of the end of RT by four experts (two radiologists and two oncologists) specialized in NPC from Hospital 1. The maximum tumor diameter (MTD) of the primary tumor and metastatic lymph node(s) were separately measured on MR images. Tumor remission was evaluated based on the change in total MTD (estimated as the sum of the MTD of the primary tumor and metastatic lymph nodes) before and after treatment. Tumor remission was divided into 4 levels: complete remission (CR); partial remission (PR); stable disease (SD); and progressive disease (PD) ². Since almost all patients with NPC will achieve good disease control after radical (chemo)radiotherapy (87.0–97.0%) ^{3–5}, CR was defined as complete remission of the disease or the absence of any residual tumor, including both the primary tumor at the nasopharynx and the metastatic cervical lymph node(s), on MR images. PR was defined as the presence of residual disease in either the primary site and/or metastatic cervical lymph node(s). SD and PD were defined according to the Response Evaluation Criteria in Solid Tumors (RECIST, version 1.1) ².

Methods 4. Follow-up

Patients were assessed every 3 months during the first 3 years, every 6 months during the next 2 years, and annually thereafter. A total of 1 723 patients with treated NPC from Hospital 1 were included between September 10, 2007, and May 20, 2021. The last follow-up date was December 31, 2021. Eighty-five patients from Hospital 2 were included between September 6, 2011, and September 4, 2020. The last follow-up date was October 30, 2021. The mean follow-up duration was 18.7 ($\pm$ standard deviation [SD]: 9.7), 19.3 ($\pm$ SD: 9.6), 18.3 ($\pm$ SD: 9.4), 44.4 ($\pm$ SD: 20.7), and 17.5 ($\pm$ SD: 10.6) months in the training, tuning, internal, external and prospective validation cohorts, respectively.

Local (regional) relapse-free survival (LRRFS) was defined as the duration between the end of the first RT course and the

diagnosis of recurrence after > 6 months to exclude partial responders to the first course of RT.

Methods 5. Clinical classification

All patients underwent H&N MR within 6 months after definitive treatment to establish a new baseline for future comparisons. CR was achieved in all eligible patients. Routine annual imaging was indicated to visualize areas inaccessible to routine clinical examination. The routine follow-up MR scans of the H&N were acquired at regular intervals within 5 years, generally at 3–24 months, depending on the oncologists' preference and patients' compliance. During the follow-up period, patients were classified as recurrent nasopharyngeal carcinoma (rNPC) cases (biopsy-, functional imaging-, or radiologically confirmed recurrence in local, regional or locoregional sites) or nonrecurrent nasopharyngeal carcinoma (non-rNPC) cases (remained in CR for 5 years starting from the completion date of the initial treatment).

In cases of rNPC, for lesions that were accessible, the diagnosis was based on the positive histopathological results from biopsy (any component with cancerous cells or tissue originating from the primary NPC). Recurrence was firstly found by physical examination, nasopharyngoscopy or imaging modality (e.g., MRI, CT). By histopathological or cytological biopsy, or pathological examination of surgery-resected lesion, the reoccurring lesion contained any component with cancerous cells or tissue originating from the primary NPC is diagnosed as recurrence. For lesions that were not accessible, the diagnoses were made based on radiologic images. For patients with available functional imaging 18F-Fluorodeoxyglucose positron emission tomography and computed tomography (18F-FDG PET/CT) imaging, the clinical diagnosis was made in consensus by two nuclear medicine physicians (each with 5 years of experience in PET/CT) using a GE Xeleris workstation. Image interpretation was based on visual evaluations. Any focus of FDG uptake greater than the surrounding background and not attributable to normal FDG biodistribution was assessed. The intensity of FDG uptake was graded using the five-point scale proposed by Ng et al ^{6,7}. The probability of recurrent tumors was graded by using a five-point scoring system (0 = no lesion, 1 = definitely benign, 2 = probably benign, 3 = probably malignant, and 4 = definitely malignant). Grade 3 and grade 4 were both considered to be positive results. The sites of the positive results were recorded. In patients with a clinical diagnosis based on MR images, pre, posttreatment and “suspected recurrence” MR scans for each patient were independently reviewed by a rotating panel of four board-certified experts (two radiologists and two oncologists) from Hospital 1 (same expert group in **Supplementary Methods 3**. Treatment response evaluation). Local recurrence was diagnosed when a focal area that showed the following was noted ⁸: (i) intermediate signal intensity on T1-weighted images; (ii) enhancement on contrast-enhanced T1-weighted images; or (iii) high signal intensity on T2-weighted images. Regional recurrence was diagnosed when the following were observed ⁹: (i) an enlarged cervical lymph node (LN) with a minimal axial diameter (MID) ≥ 10 mm in the largest plane; (ii) an enlarged lateral retropharyngeal lymph node with an MID ≥ 5 mm or any enlarged node in the median retropharyngeal group; (iii) an LN of any size with central necrosis or a contrast-enhanced rim; (iv) nodal grouping, the presence of ≥ 3 contiguous and confluent LNs, each with an MID 8–10 mm; or (v) an LN of any size with extracapsular spread. All enhanced lesions on MR images, irrespective of patient age, were considered tumor recurrence until proven otherwise. Fibrosis was diagnosed with confidence only if there is no contrast enhancement coupled with low signal intensity on T2-weighted images.

In cases of non-rNPC, CR was defined as complete resolution of the disease or absence of any residual tumor in both the primary and nodal sites on posttreatment MR scans, also including stable posttreatment changes seen on later imaging within a 5-year follow-up period.

Methods 6. Imaging scanning protocol

The follow-up MR examinations were performed on a 1.0-T MRI unit ([Panorama HFO; PHILIPS, Netherlands]), or 1.5-T MRI unit ([Achieva; PHILIPS, Netherlands]; [Magnetom Aera; SIEMENS, Germany]; [Magnetom Espree; SIEMENS, Germany]; [uMR 560; UNITED IMAGING, China]) or 3.0-T MRI unit ([Achieva 3.0T; PHILIPS, Netherlands]; [Discovery MR750w; GE, USA]; [Ingenia CX; PHILIPS, Netherlands]; [Magnetom Prisma; SIEMENS, Germany]; [Magnetom Trio Tim; SIEMENS, Germany]; [Pioneer; GE, USA]; [Premier; GE, USA]; [Signa Architect; GE, USA]; [uMR 780; UNITED IMAGING, China]; [uMR 790; UNITED IMAGING, China]). Scanning was performed with a head-neck combined coil. The scanning range was from the suprasellar cistern to the inferior margin of the sternal end of the clavicle. T1- and T2-weighted fast spin-echo images in the axial direction were obtained before the injection of contrast material. Enhanced T1-weighted axial sequences were performed 40 seconds after an intravenous Gd-DTPA (Magnevist; Bayer Schering Pharma AG, Germany) injection at a dose of 0.1 mmol/kg. The vendor-

specific acquisition parameters of each device for H&N MR scans are summarized in **Supplementary Table 9**. All H&N MR images were stored in Digital Imaging and Communications in Medicine (DICOM) format in the imaging databases of two hospitals.

When there is difficulty or uncertainty in locating recurrent lesions on MR images, the experts who labeled the scans can refer to the result of 18F-FDG PET/CT examination with 2 nuclear medicine specialists. 18F-FDG PET/CT scans were performed using a Discovery ST-16 (GE Medical Systems, Milwaukee, WI, USA) in two hospitals. Imaging was performed using a combined PET/CT scanner according to PET/CT tumor imaging guidelines. First, the patients fasted for 6 hours prior to the PET/CT scan. Iohexol (Omnipaque; GE Healthcare) was used as the oral contrast agent with 500 ml water. The blood glucose levels obtained immediately before 18F-FDG injection were maintained at <11.1 mmol/L in all patients. Then, imaging was performed 40–60 minutes after an injection of 3.7 MBq/kg of body weight (0.1 mCi/kg) of 18F-FDG. Next, a low-dose multislice CT scan was acquired from the skull base to the upper thigh base using a 16-slice multidetector scanner (parameters: 180–250 mA, 140 kV, pitch 1.375 mm, and slice thickness 3.75 mm) while the patient performed shallow breathing. The patients' arms were placed up by their heads to avoid artefacts from the liver FDG signal due to beam-hardening effects. The standard uptake value (SUV) was calculated using the body weight. A standard whole-body PET scan was performed in 2D mode with an acquisition time of 3 minutes per bed position (six-eight bed positions) covering the same field as the CT scan. The acquired data were reconstructed using an iterative algorithm, and CT images without contrast enhancement were acquired for attenuation correction. Finally, the acquisition data were transferred to a workstation (AW Server 2.0; GE Healthcare) for processing and interpretation. The transaxial, coronal and sagittal CT, PET and coregistered PET/CT images were evaluated by two experienced nuclear medicine physicians. Differences were resolved by consensus.

Methods 7. Scan labelling

All patients' follow-up H&N MR scans within the follow-up period were evaluated at the individual level by a rotating panel of four board-certified experts (two radiologists and two oncologists) from Hospital 1 (same expert group in **Supplementary Methods 3**) following the labelling instructions listed in **Supplementary Table 10**. Each expert had at least 5 years of experience specializing in NPC. The experts were free to indicate any suspected recurrence lesion on the MR images. A scan was considered nonrecurrent based on the following findings: 1) no evidence of new lesions on MR images; 2) histopathologically confirmed inflammation or adenoid hyperplasia; and 3) stable disease after radiotherapy with at least one-year follow-up. The exclusion criteria were as follows: 1) histopathologically confirmed secondary tumor or second primary tumor; 2) single examination without previous or subsequent follow-up MR scans; or 3) divergent diagnoses with failure to reach a consensus. A follow-up H&N MR scan was labelled positive for recurrence based on the following findings: 1) histopathologically confirmed recurrence; 2) newly developed or enlarged lesion with enhancement located in the nasopharynx or adjacent structures compared to that shown on previous imaging; 3) newly developed or enlarged lymph nodes with metastatic signs (e.g., shorter diameter ≥ 15 mm, necrotic, extracapsular spread, confluent) compared to that shown on previous imaging; and 4) newly developed or enlarged lesions located in the nasopharynx or adjacent structures or lymph nodes with hypermetabolism on functional imaging (e.g., positron emission tomography). Once a 'suspected' radiologically visible lesion was confirmed to be recurrence by clinical records, a labelling procedure was followed to determine the scan with the earliest visible recurrence scan (**Supplementary Fig. 15 and Supplementary Table 10**). Based on the typical and apparent recurrent lesions on the MR images at the diagnosis of recurrence, the MR images prior to that date were traced backtracks by the same group of experts to find the scan that showed the first signs of recurrence. Experts could consider all known clinical information to make a final decision. Disagreements about the scans were arbitrated and the final decision was decided by a senior oncologist and a senior radiologist independently of the original four labellers, but with knowledge of their labels. The image labels were finalized only when the four experts from the same group reached a consensus.

Because other imaging modalities often used in clinical practice, such as nasopharyngeal endoscopy, H&N computed tomography, and whole-body positron emission tomography, were unavailable (as these are not routinely performed during surveillance), we describe the MR-derived early recurrence label as a silver standard. From this process, the 'true' recurrence scans of 402 (44.5%) rNPC patients were readjusted (**Supplementary Fig. 1–3**). The initial dataset of patients with confirmed recurrence consisted of 9 309 (91.1%) conservative scans and 903 (8.9%) diagnostic recurrence scans. The final dataset, after labelling, consisted of 8 793 (86.1%) conservative scans, 402 (3.9%) early recurrence scans, 114 (1.1%) visible recurrence scans (between early recurrence and diagnostic recurrence), and 903 (8.8%) confirmed recurrence scans (**Supplementary Fig. 1–3**). When the early recurrence scan

identified by experts and the diagnostic recurrence scan overlapped, the label was defined as the scan with the earliest visible recurrence.

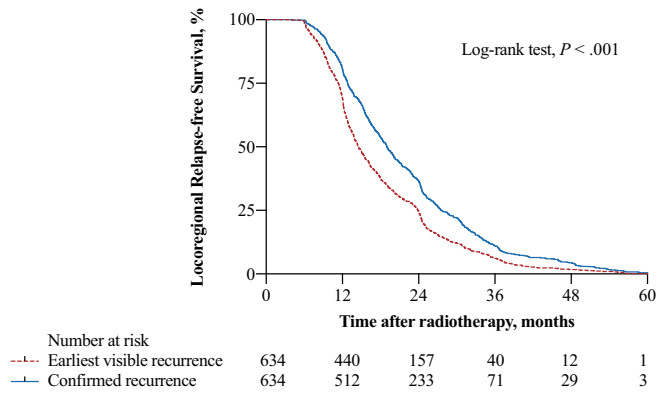
The follow-up H&N MR scans and basic clinicopathological information (sex, age, recurrent tumor type, recurrent stage, diagnosis of recurrence) of each patient were independently retrieved from the Hospital Information System (HIS) in two centers and processed in Hospital 1.

Supplementary References

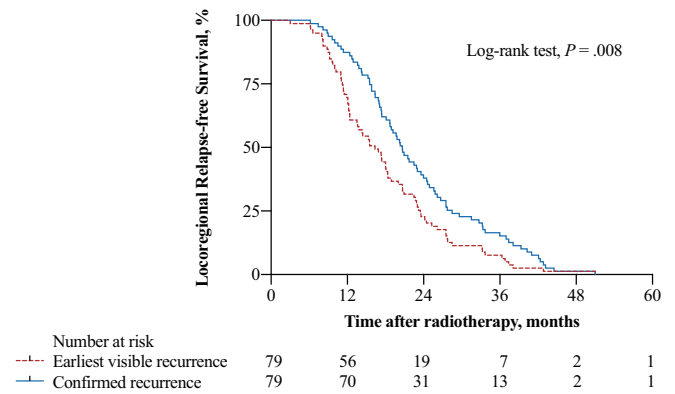
1. Amin, M. B. *et al.* The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more ‘personalized’ approach to cancer staging. *CA. Cancer J. Clin.* **67**, 93–99 (2017).
2. Therasse, P. *et al.* New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *JNCI J. Natl. Cancer Inst.* **92**, 205–216 (2000).
3. Lee, A. W. M. *et al.* Evolution of treatment for nasopharyngeal cancer – Success and setback in the intensity-modulated radiotherapy era. *Radiother. Oncol.* **110**, 377–384 (2014).
4. Zhang, M.-X. *et al.* Intensity-modulated radiotherapy prolongs the survival of patients with nasopharyngeal carcinoma compared with conventional two-dimensional radiotherapy: A 10-year experience with a large cohort and long follow-up. *Eur. J. Cancer* **51**, 2587–2595 (2015).
5. Mao, Y.-P. *et al.* Prognostic factors and failure patterns in non-metastatic nasopharyngeal carcinoma after intensity-modulated radiotherapy. *Chin. J. Cancer* **35**, 103 (2016).
6. Ng, S.-H. *et al.* Nodal metastases of nasopharyngeal carcinoma: patterns of disease on MRI and FDG PET. *Eur. J. Nucl. Med. Mol. Imaging* **31**, 1073–1080 (2004).
7. Ng, S.-H. *et al.* Comprehensive imaging of residual/ recurrent nasopharyngeal carcinoma using whole-body MRI at 3 T compared with FDG-PET-CT. *Eur. Radiol.* **20**, 2229–2240 (2010).
8. Chong, V. F. & Fan, Y. F. Detection of recurrent nasopharyngeal carcinoma: MR imaging versus CT. *Radiology* **202**, 463–470 (1997).
9. Li, W.-F. *et al.* Proposed lymph node staging system using the International Consensus Guidelines for lymph node levels is predictive for nasopharyngeal carcinoma patients from endemic areas treated with intensity modulated radiation therapy. *Int. J. Radiat. Oncol. Biol. Phys.* **86**, 249–256 (2013).

Part II. Supplementary Figures

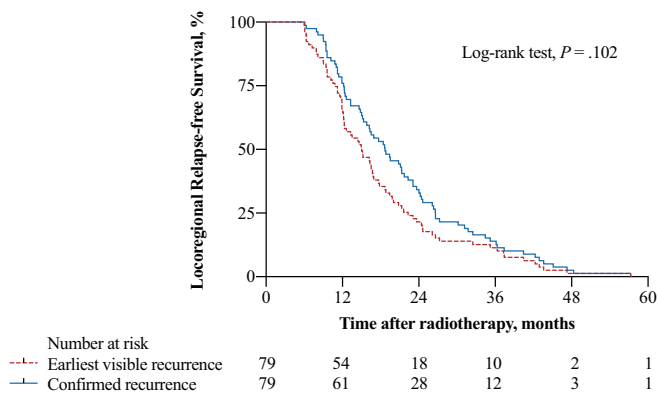
Training



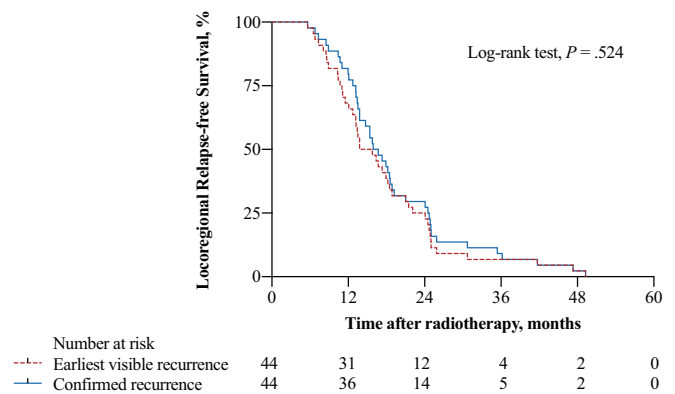
Tuning



Internal Validation



External Validation



Prospective Validation

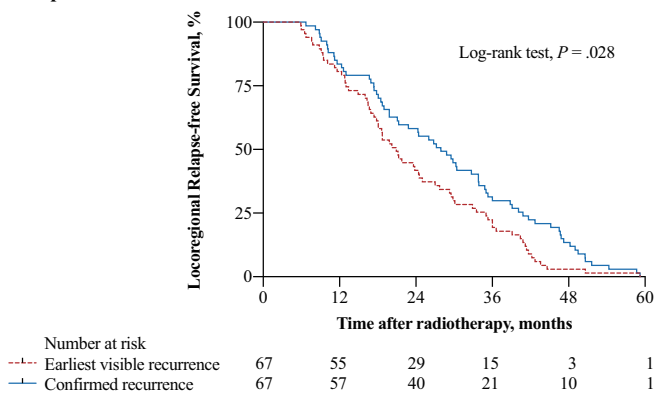


Fig. 1 Kaplan-Meier curves for locoregional relapse-free survival at the earliest visible recurrent nasopharyngeal carcinoma lesion and when recurrent nasopharyngeal carcinoma was confirmed. *Actuarial survival rates were calculated using the Kaplan-Meier method and compared using the log-rank test.

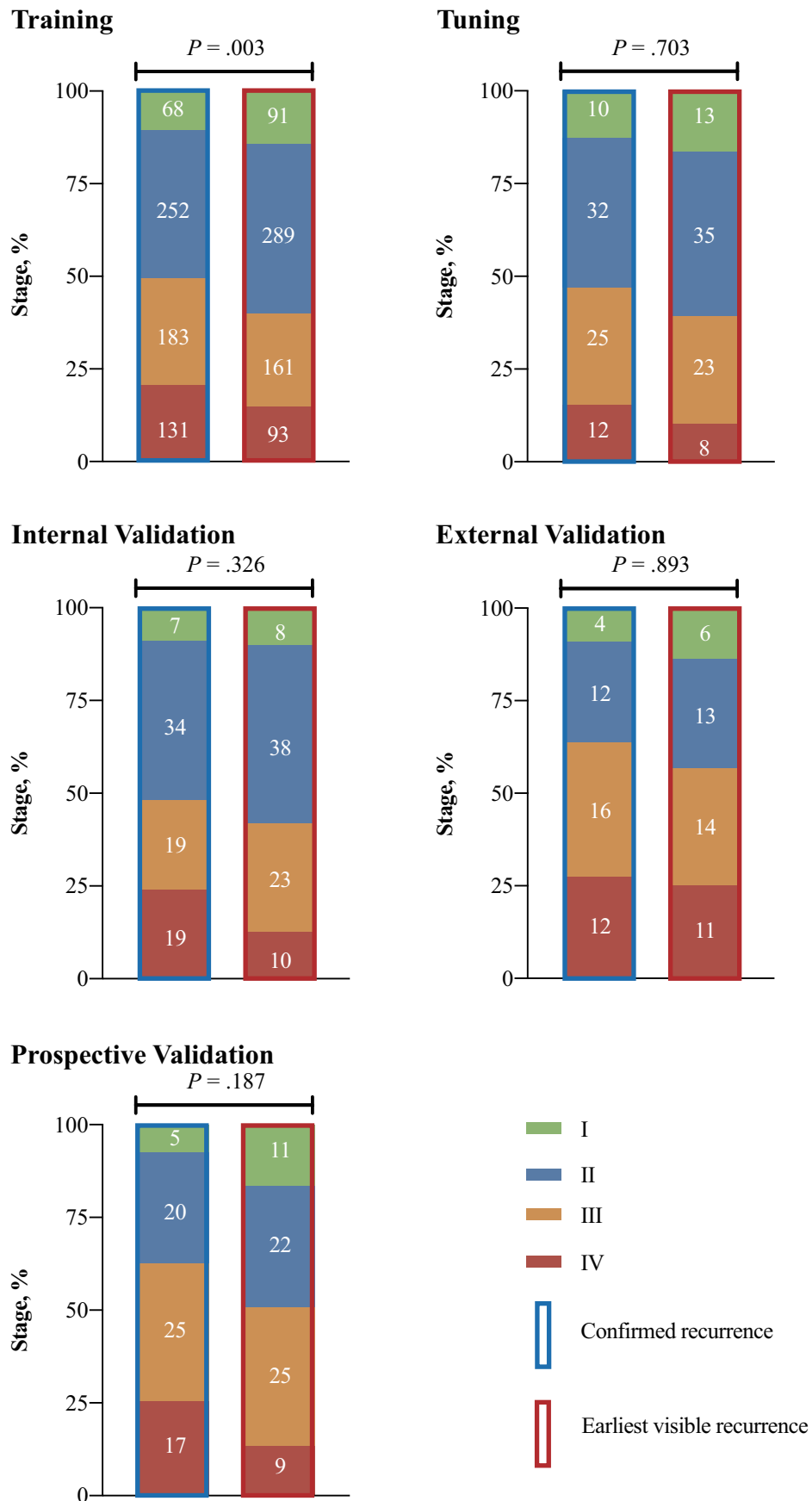
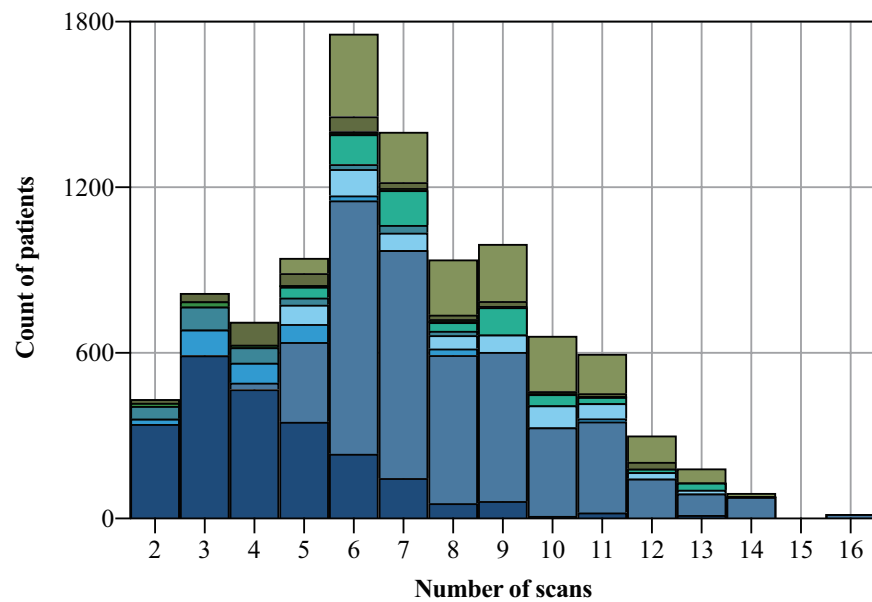


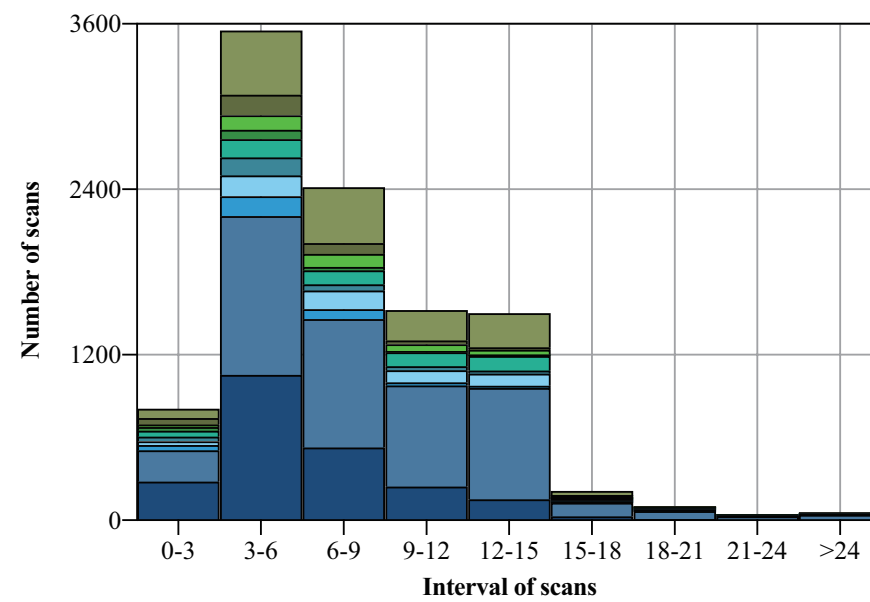
Fig. 2 Distributions of the clinical staging of recurrence at the earliest visible recurrent nasopharyngeal carcinoma lesion and when recurrent nasopharyngeal carcinoma was confirmed. Data are presented as the number

in the histogram. *The 8th edition of the American Joint Committee on Cancer (AJCC) guidelines was used for recurrent tumor and metastatic lymph node(s) staging. **Pearson's chi-square test or Fisher's exact test were used to assess categorical variables between cohorts.

A



B



C

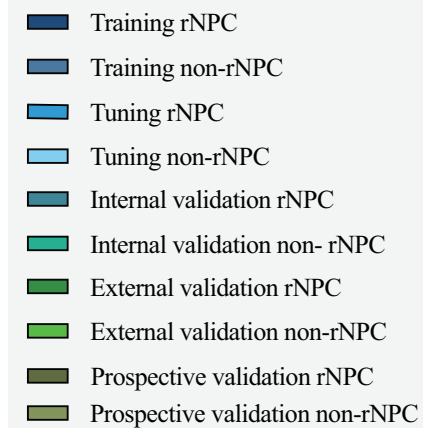
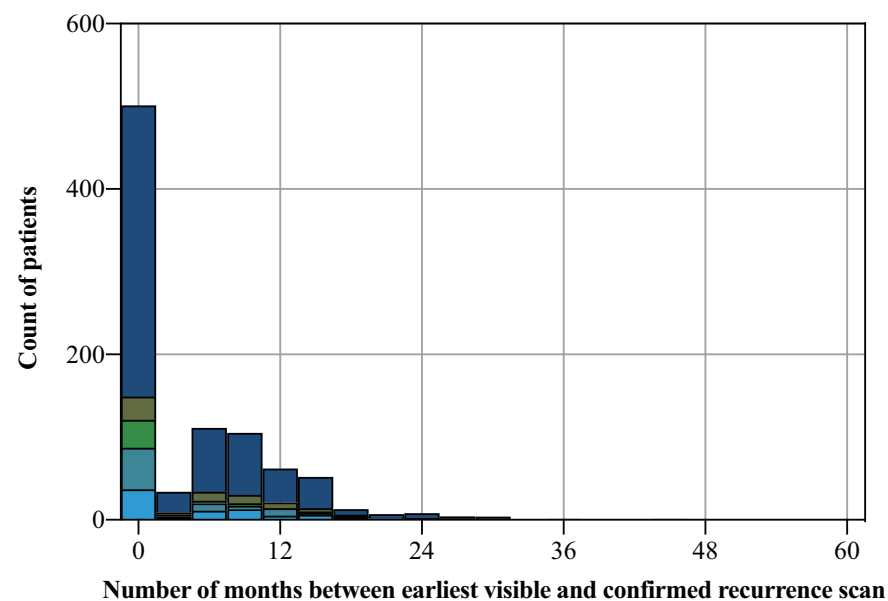


Fig. 3 Dataset statistics. (A) Histogram of scans per survivor according to disease status in the training, tuning, internal validation, external validation, and prospective validation cohorts.

(B) Histogram of the interval between follow-up scans per survivor according to disease status in the training, tuning, internal validation, external validation, and prospective validation cohorts.

(C) Histogram of the time interval between scans the earliest visible recurrence and confirmed recurrence for recurrent nasopharyngeal carcinoma in the training, tuning, internal validation, external validation and prospective validation cohorts. rNPC, recurrent nasopharyngeal carcinoma; non-rNPC, non-recurrent nasopharyngeal carcinoma.

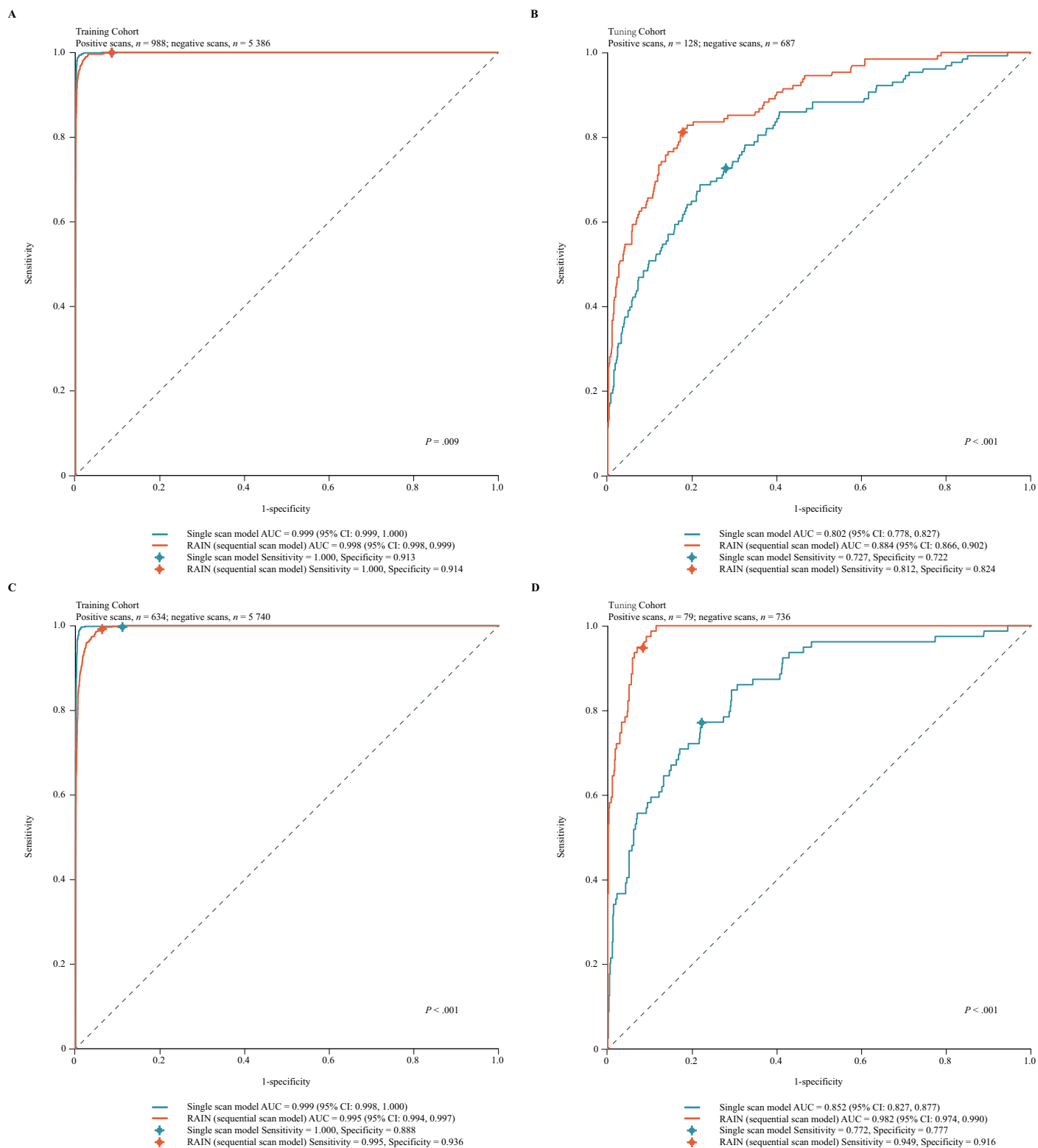


Fig. 4 Receiver operating characteristic curves of the RAIN model (sequential scan model) vs single scan model for discriminating follow-up H&N MR scans in the training and tuning cohorts. Performance of the RAIN model in identifying scans with recurrence in the training (A) and tuning (B) cohorts. Performance of the RAIN model in identifying scans with confirmed recurrence in the training (C) and tuning (D) cohorts. H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; AUC, area under the receiver operating characteristic

curve; CI, confidence interval.

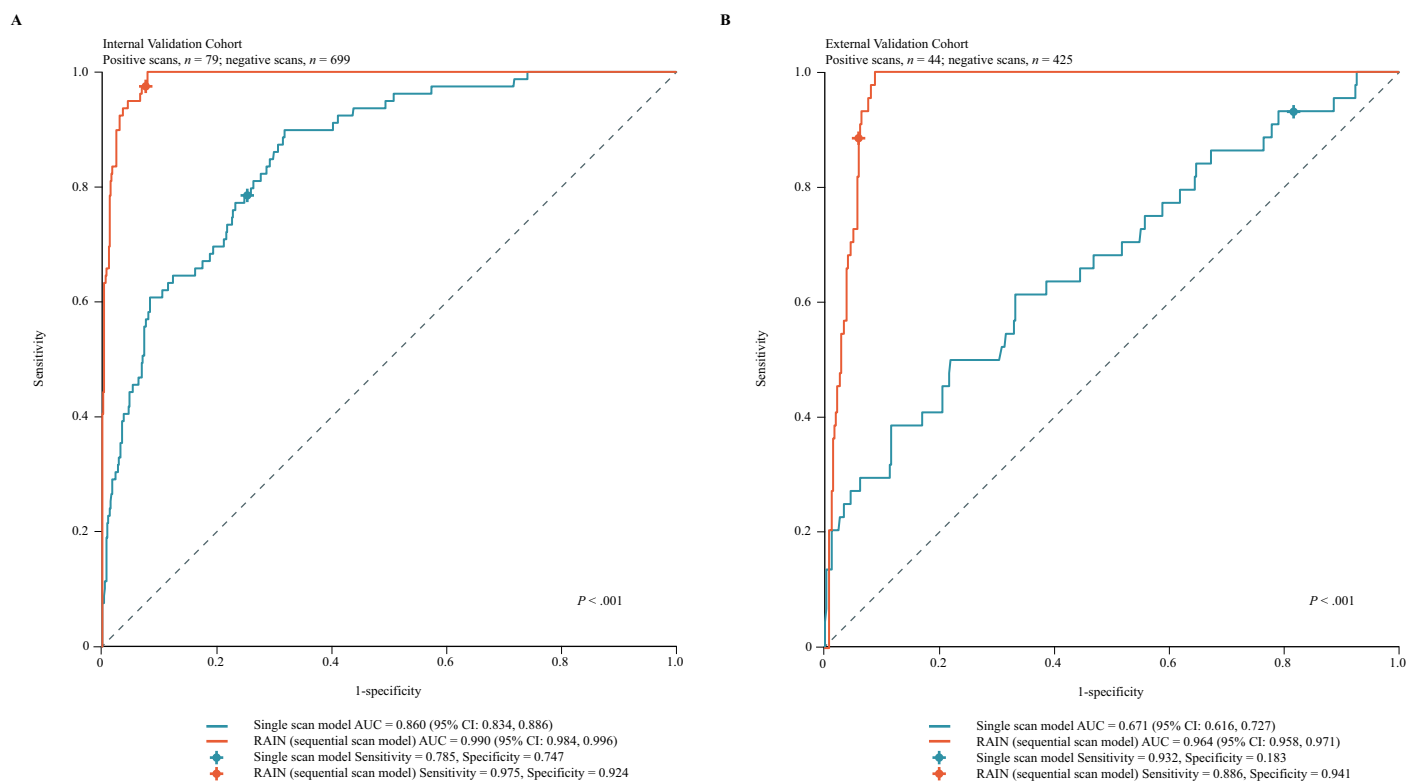


Fig. 5 Receiver operating characteristic curves of the RAIN model (sequential scan model) vs single scan model

for discriminating follow-up H&N MR scans in the internal and external validation cohorts. Performance of the

RAIN model in identifying scans with confirmed recurrence in the internal (**A**) and external (**B**) validation cohorts. H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; AUC, area under the receiver operating characteristic curve; CI, confidence interval.

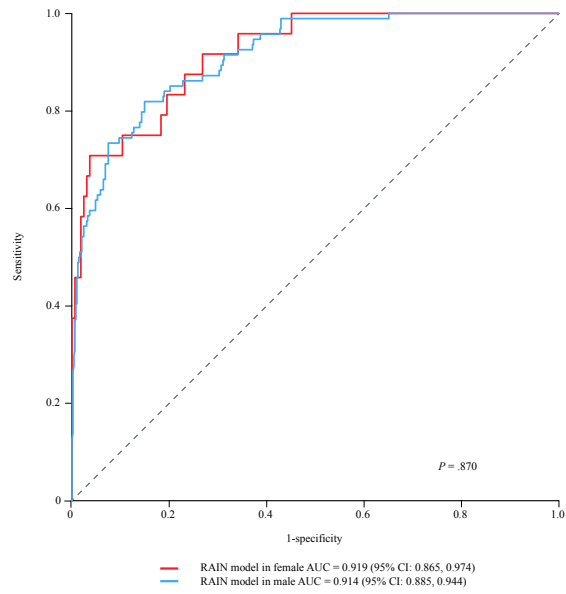
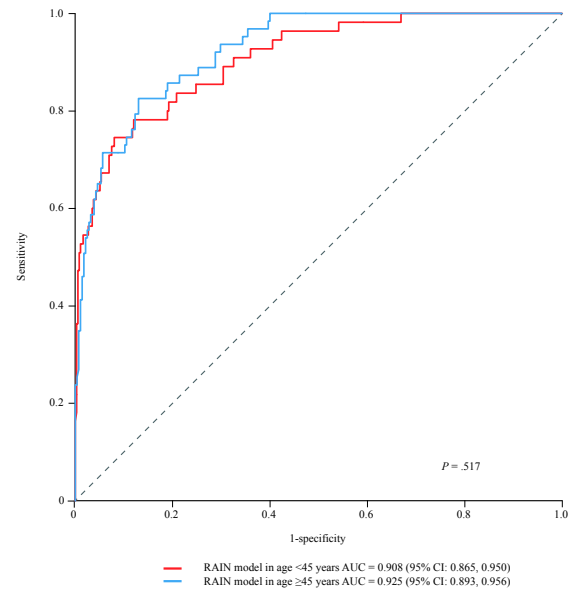
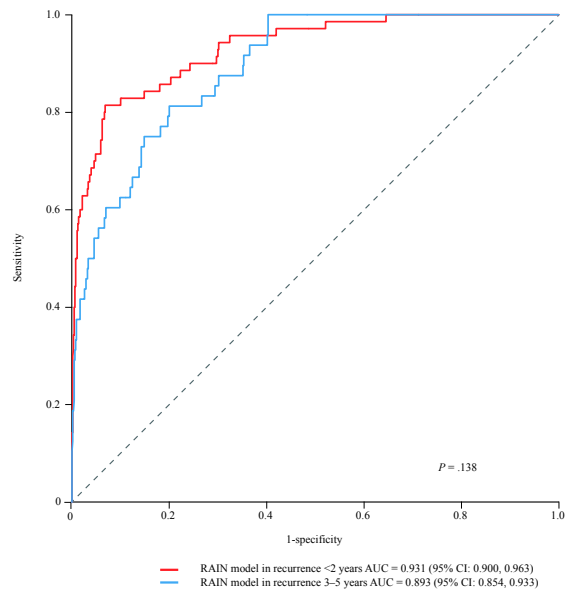
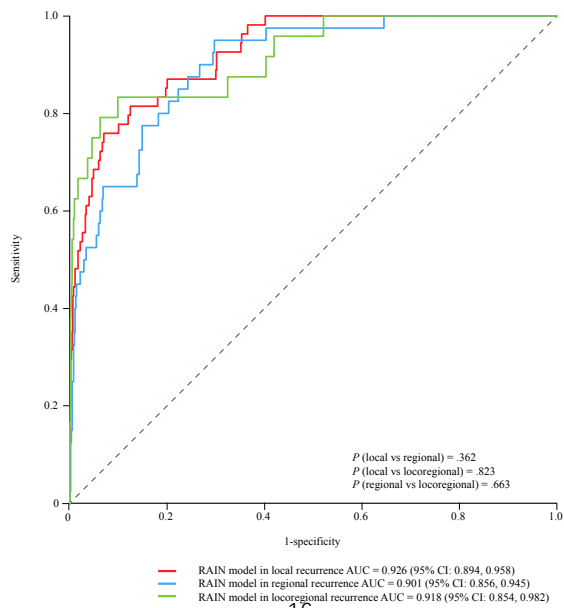
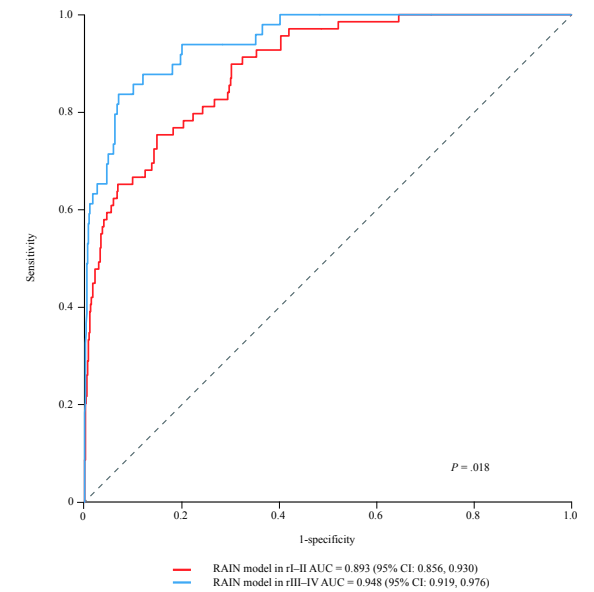
A**B****C****D****E**

Fig. 6 Receiver operating characteristic curves of the RAIN model (sequential scan model) in discriminating follow-up H&N MR scans in the internal validation cohort stratified by different clinical factors. H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; AUC, area under the receiver operating characteristic curve; CI, confidence interval.

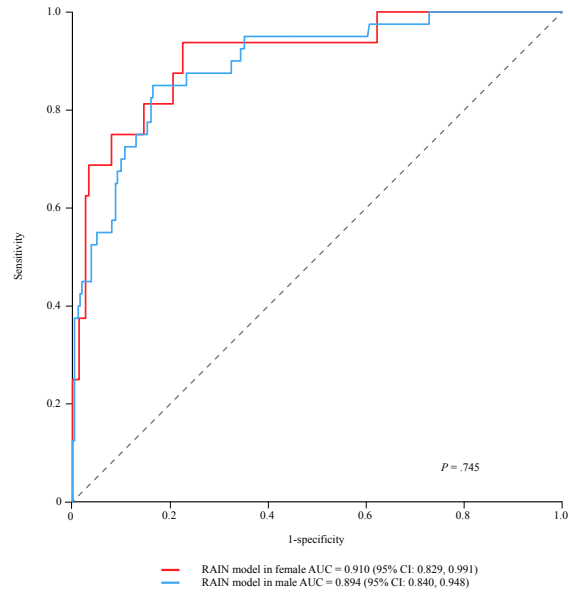
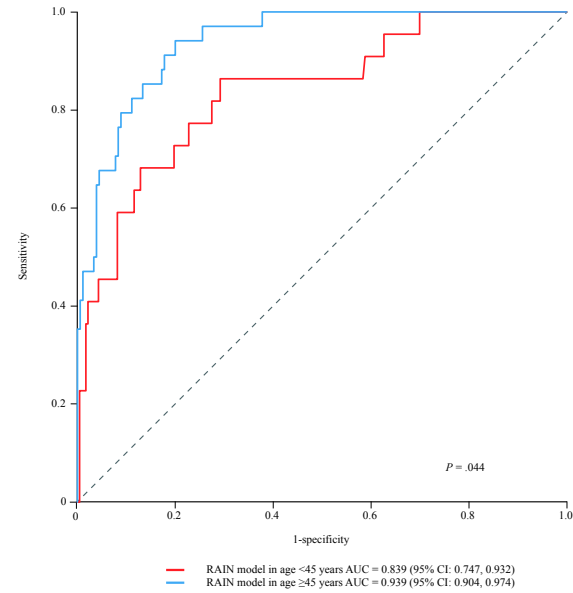
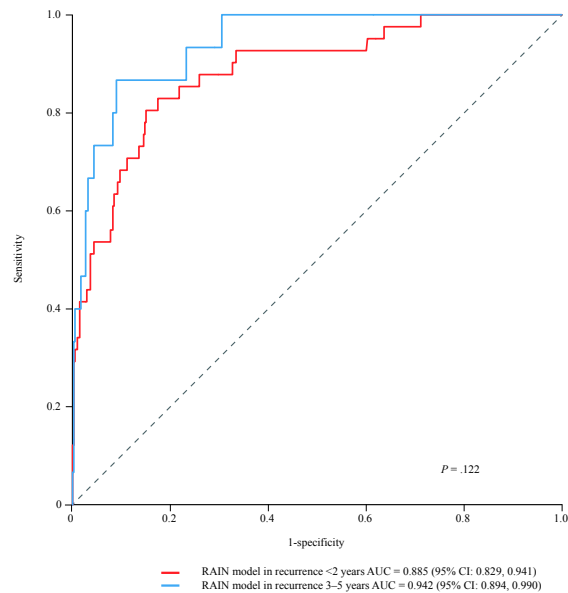
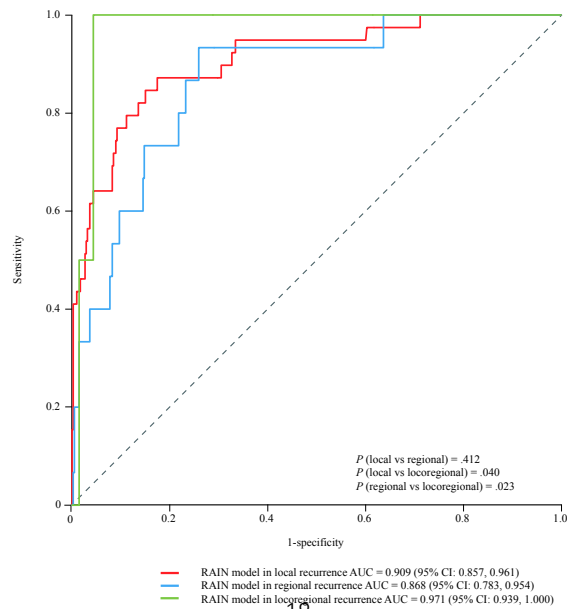
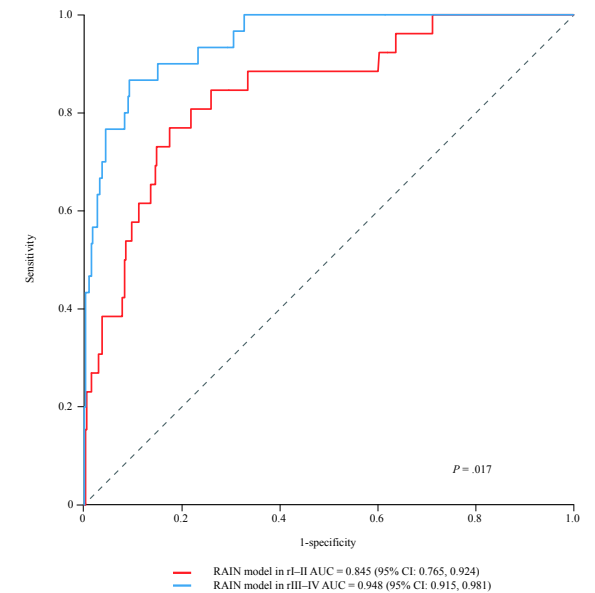
A**B****C****D****E**

Fig. 7 Receiver operating characteristic curves of the RAIN model (sequential scan model) in discriminating follow-up H&N MR scans in the external validation cohort stratified by different clinical factors. H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; AUC, area under the receiver operating characteristic curve; CI, confidence interval.

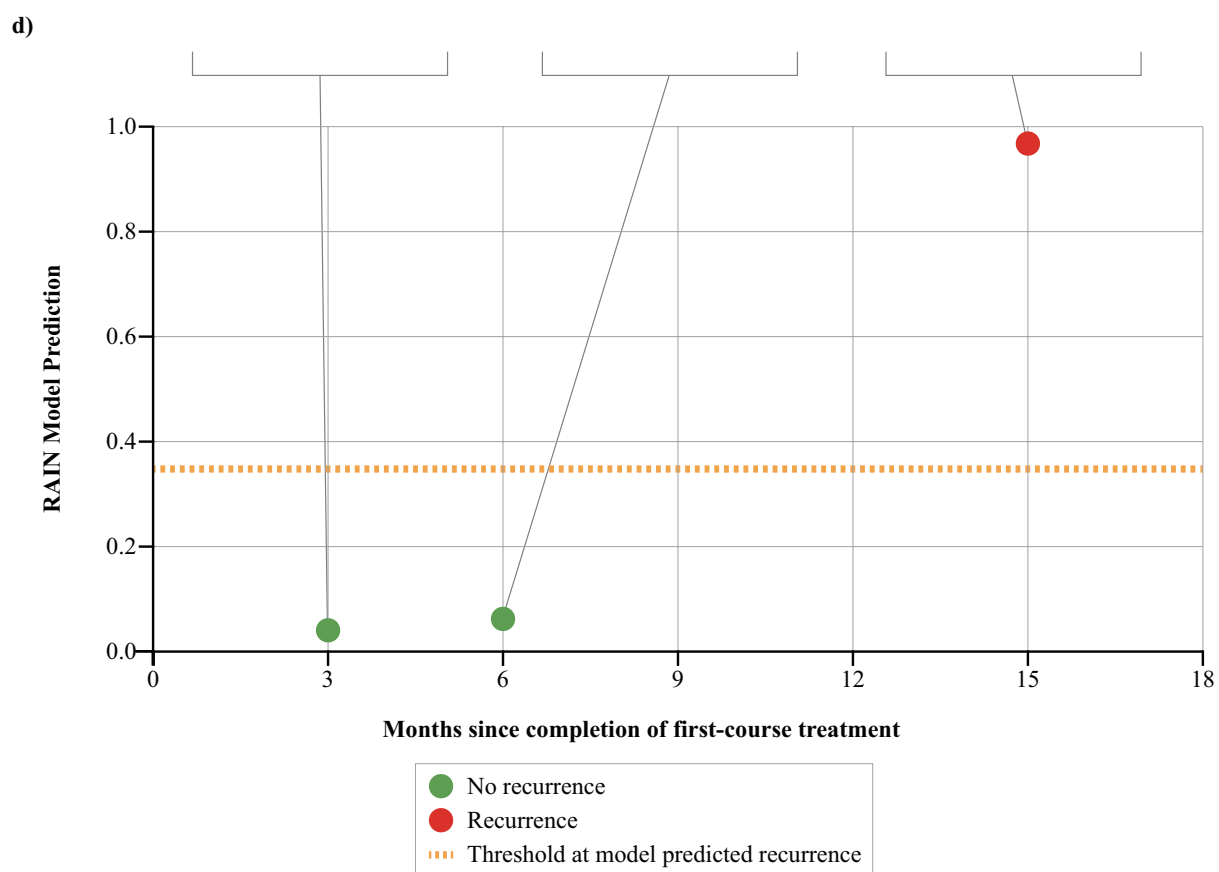
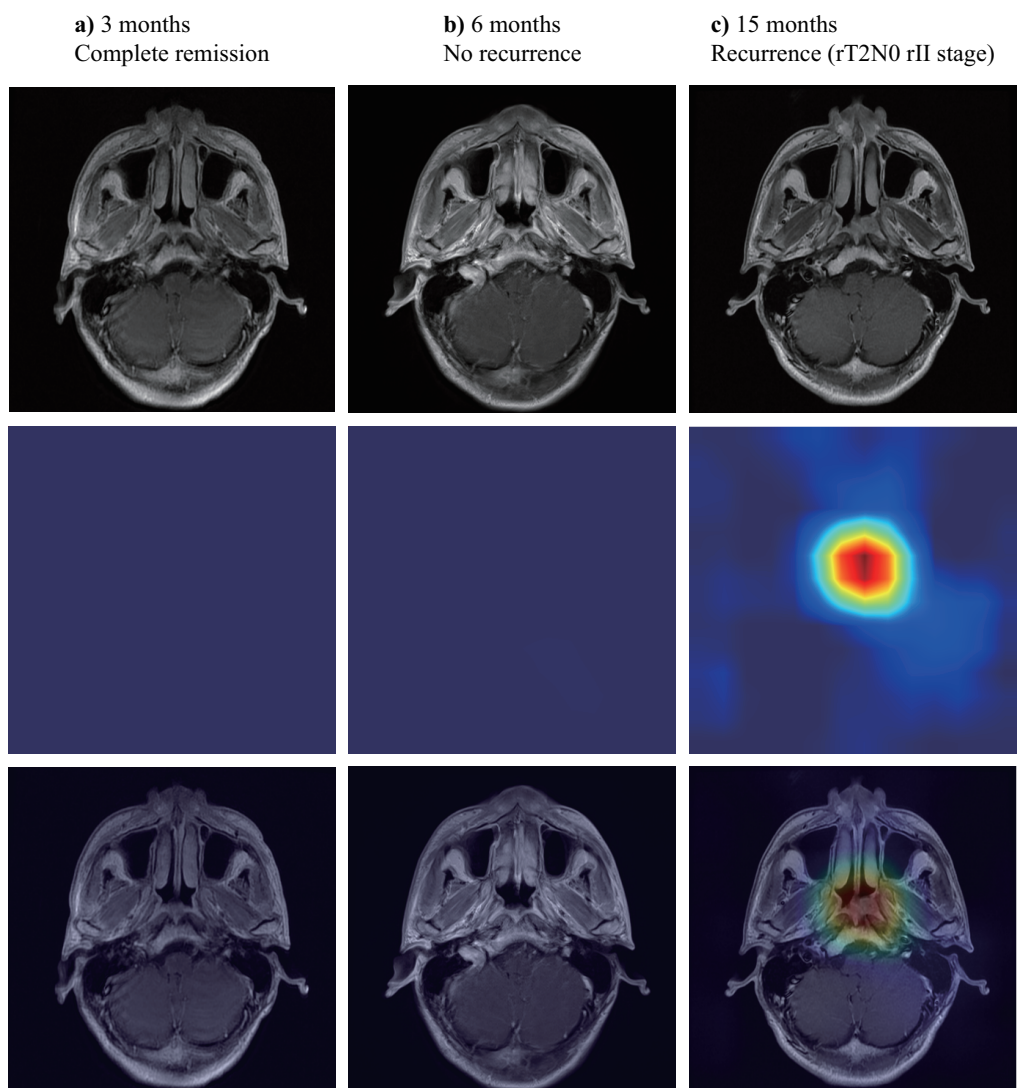


Fig. 8 A successful case of a true positive in a 51-year-old male who presented with recurrence in the left fossa of rosenmüller. The patient was diagnosed with nasopharyngeal carcinoma (T3N2M0 III stage [AJCC Staging System 8th version]) at the initial presentation. He received radical concurrent chemoradiotherapy. Follow-up H&N MR scans were routinely performed at every visit. **a)** The patient achieved complete remission of the primary tumor 3 months posttreatment. No highlighted region was observed in the saliency map below. **b)** Six months after the first course of treatment, the treated survivor remained in complete remission from disease. **c)** However, 15 months after treatment, a neoplasm appears in his left fossa of Rosenmüller. At this point, the RAIN model predicted that this patient had 'recurrence'. Additionally, an area that the RAIN model recognized was highlighted in the corresponding saliency map below. At this point, the histopathological result of the neoplasm indicated a diagnosis of recurrence. **d)** Risk prediction graph showing the predicted probability of recurrence by the RAIN model in this case. T, tumor; N, lymph node(s); M, metastasis; AJCC, American Joint Committee on Cancer; H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

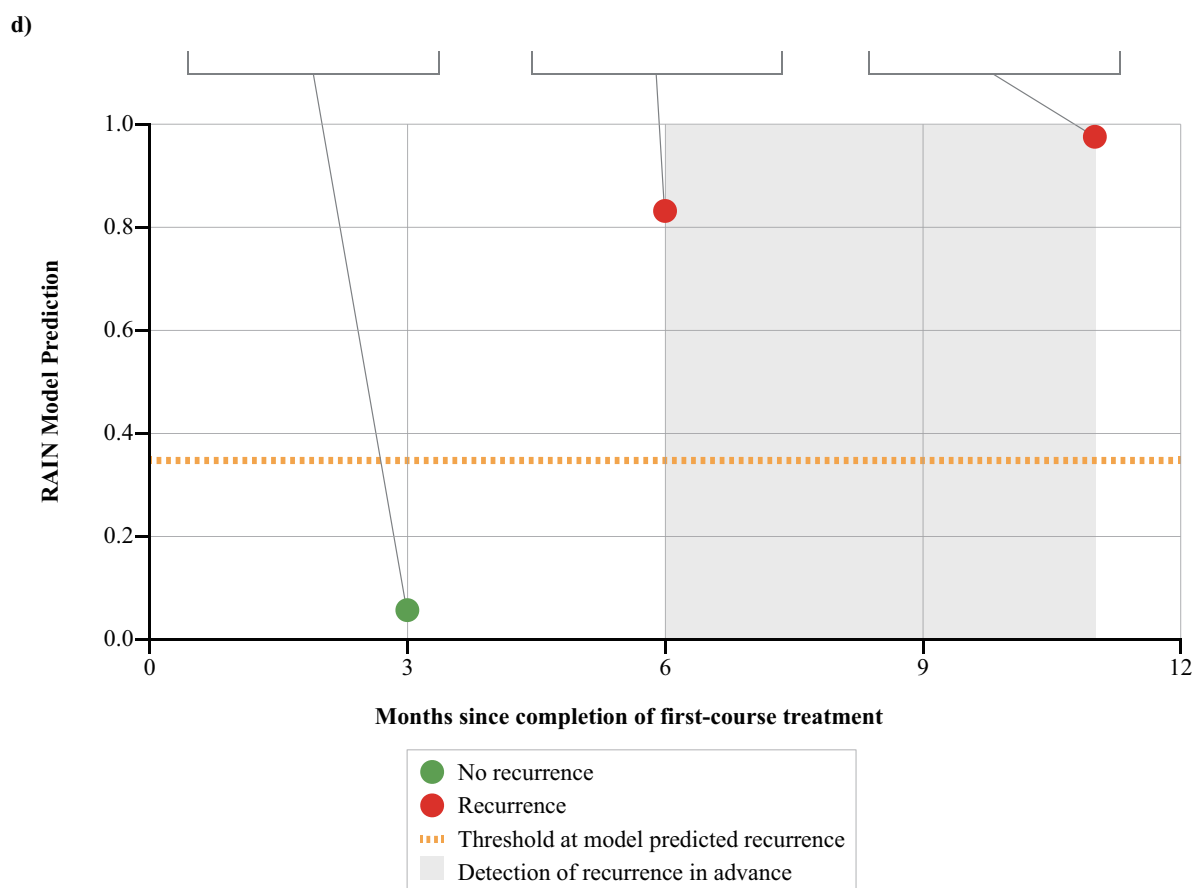
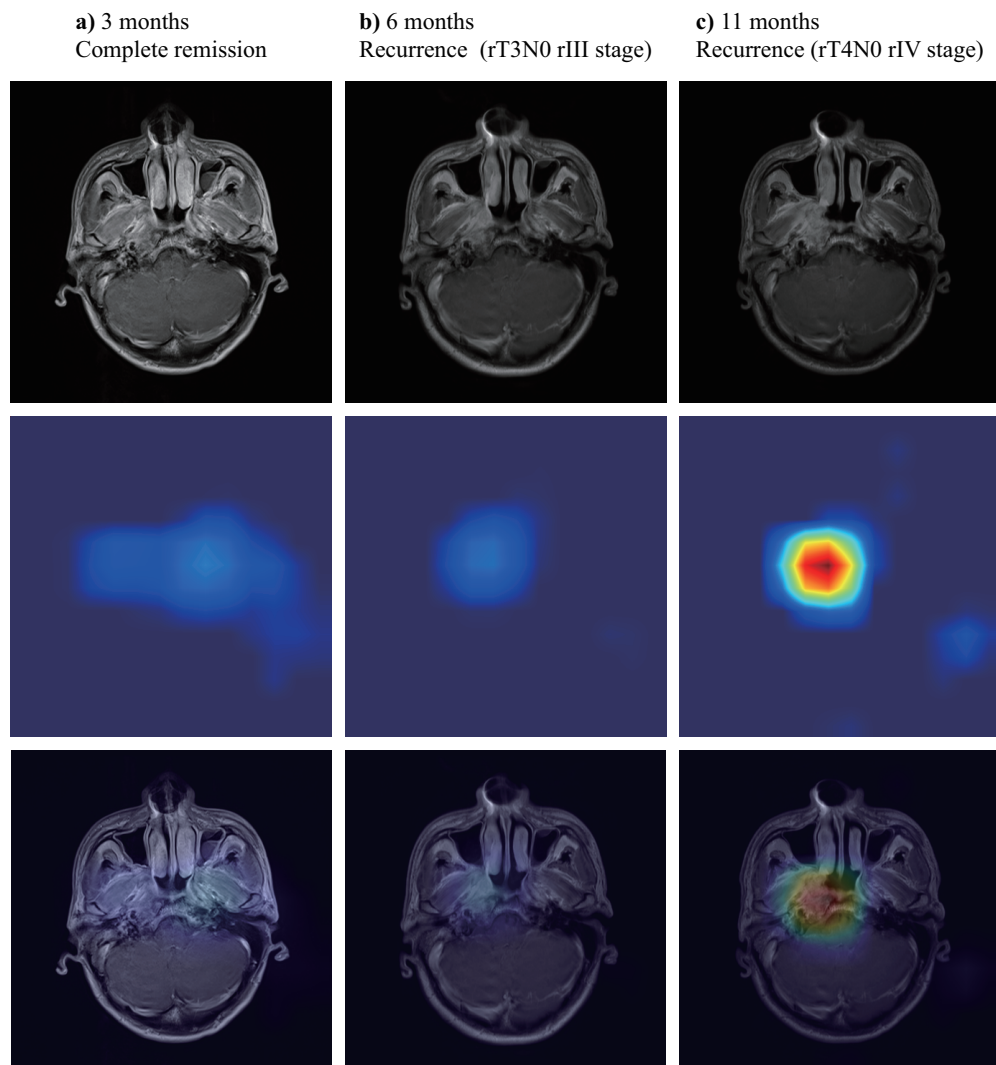


Fig. 9 A successful case of a true positive in a 33-year-old male who presented with recurrence in the right parietal wall and right lateral wall of the nasopharynx involving the right palatine fossa, medial pterygoid, musculus longus capitis, foramen lacerum and cavernous sinus. The patient was diagnosed with nasopharyngeal carcinoma (T3N2M0 III stage [AJCC Staging System 8th version]) at the initial presentation. He received radical concurrent chemoradiotherapy. Follow-up H&N MR scans were routinely performed at every visit. **a)** The patient achieved complete remission of the primary tumor 3 months posttreatment. No highlighted region was observed in the saliency map below. **b)** Six months after the first course of treatment, the treated survivor was misdiagnosed as having ‘no recurrence’. At this point, the RAIN model predicted that this patient has ‘recurrence’. The MR scan showed that the lesions at the base of the right pterygoid and parapharyngeal were enlarged with an enhanced signal. At this point, the saliency map also suggested the suspicious area. However, the patient was misdiagnosed. He received no further intervention and continued routine follow-up. **c)** Five months after the previous scan, the ‘missed’ lesions progressed, enlarging in size and involving more adjacent structures. At this point, the RAIN model predicted that this patient has ‘recurrence’. Additionally, an area that the RAIN model recognized was highlighted in the corresponding saliency map below. At this point, the histopathological result of a sample from the right lateral wall of the nasopharynx indicated a diagnosis of recurrence. **d)** Risk prediction graph showing the risk of recurrence rising above the threshold at the second scan prior to progression. T, tumor; N, lymph node(s); M, metastasis; AJCC, American Joint Committee on Cancer; H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

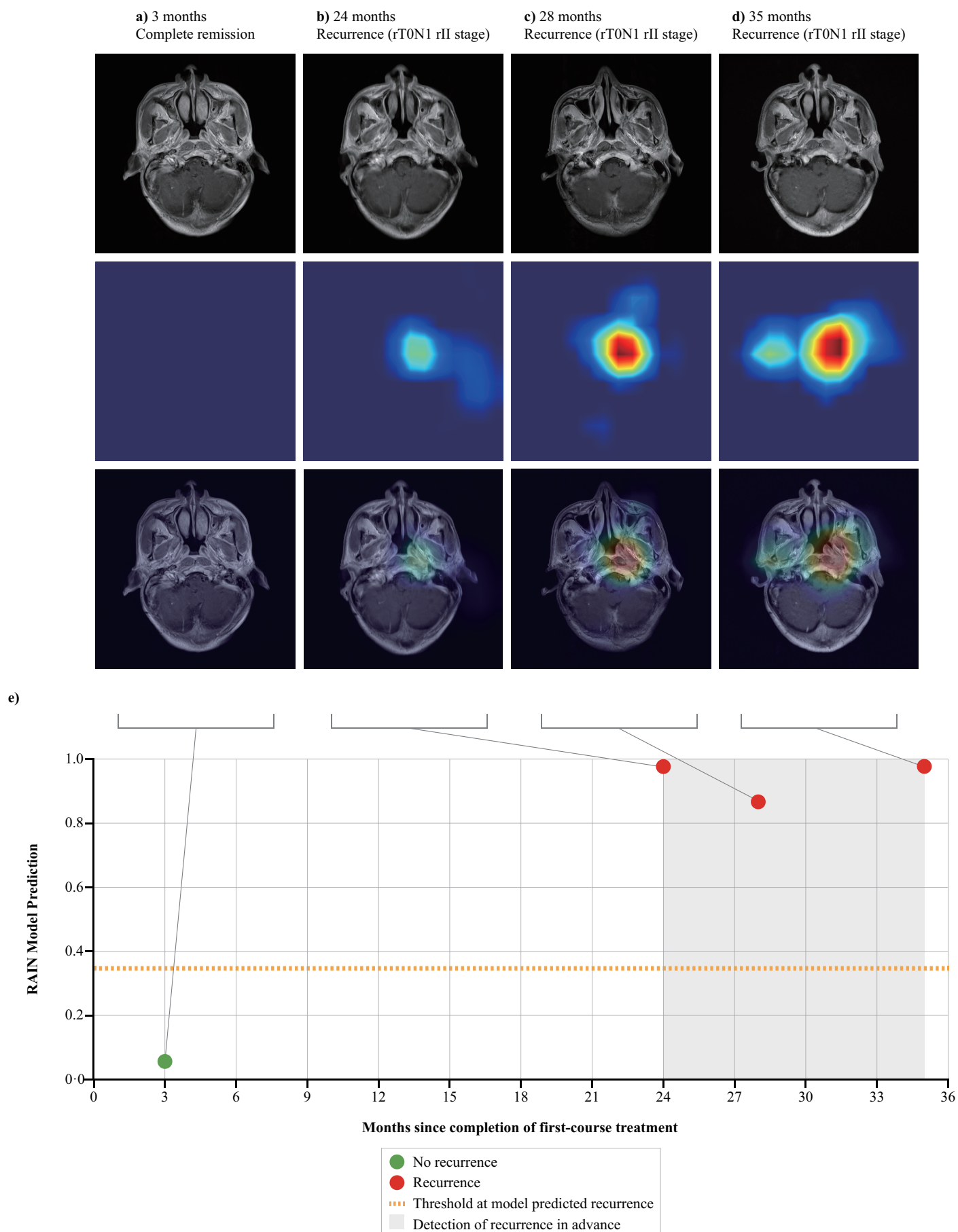


Fig. 10 A successful case of a true positive in a 46-year-old male who presented with recurrence in the left retropharyngeal lymph node. The patient was diagnosed with nasopharyngeal carcinoma (T3N2M0 III stage [AJCC Staging

System 8th version]) at the initial presentation. He received radical concurrent chemoradiotherapy. Follow-up H&N MR scans were routinely performed at every visit. **a)** The patient achieved complete remission of the primary tumor 3 months posttreatment. No highlighted region was observed in the saliency map below. **b)** Twenty-four months after the first course of treatment, the MR scan showed an enlarged left retropharyngeal lymph node. At this point, the RAIN model predicted that this MR scan showed ‘recurrence’. The saliency map also suggested the suspicious area. A few cells with heterogeneous nuclei were found in the pathological result of the cytological sample from the left retropharyngeal lymph node. However, the patient was misdiagnosed. He received no further treatment and continued routine follow-up. **c)** Four months after the previous scan, the ‘missed’ lymph node progressed, enlarging in size and showing extracapsular spread. At this point, the RAIN model predicted that this patient had ‘recurrence’ again. Additionally, an area that the RAIN model recognized was highlighted in the corresponding saliency map below. At this point, the radiologist and clinician considered these findings posttreatment changes. Again, the patient was not diagnosed with recurrence. **d)** Seven months after the previous scan, the ‘missed’ lymph node progressed, demonstrating involvement of the left levator palatini muscle, musculus longus capitis and carotid sheath. At this point, the histopathological result of the cytological sample from the left retropharyngeal lymph node indicated a diagnosis of recurrence. **e)** Risk prediction graph showing that the risk of recurrence starts to rise above the threshold at the second scan prior to progression. T, tumor; N, lymph node(s); M, metastasis; AJCC, American Joint Committee on Cancer; H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

a) 3 months
Complete remission

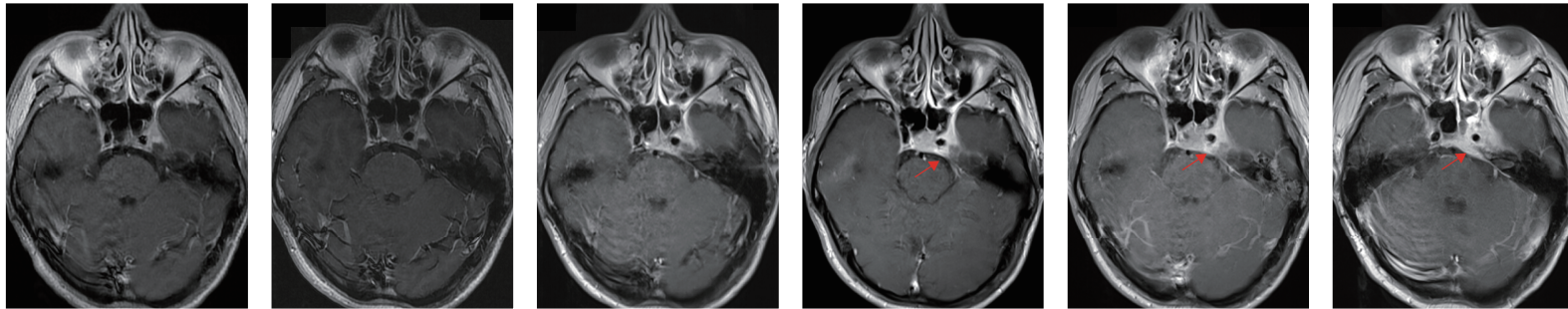
b) 6 months
No recurrence

c) 12 months
No recurrence

d) 15 months
Recurrence (rT4N0 rIV stage)

e) 19 months
Recurrence (rT4N0 rIV stage)

f) 26 months
Recurrence (rT4N0 rIV stage)



g)

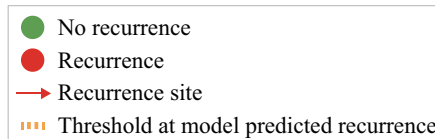
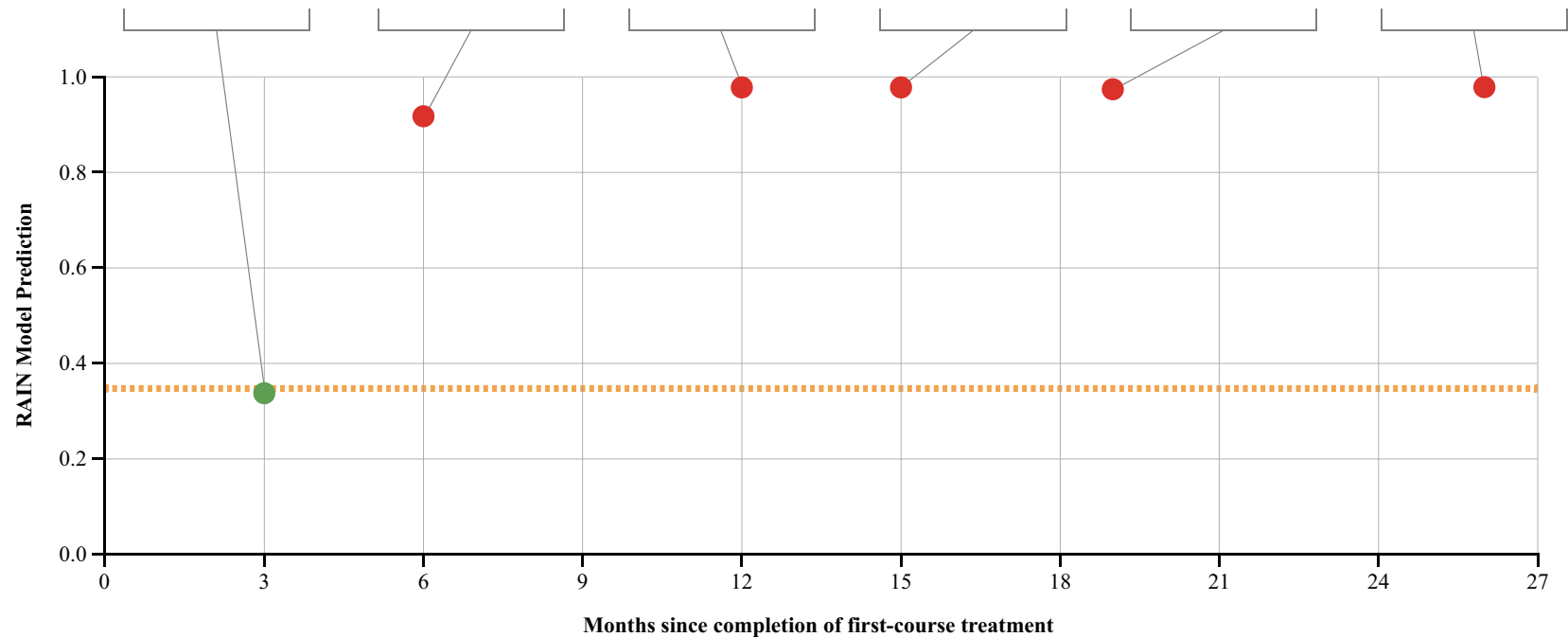
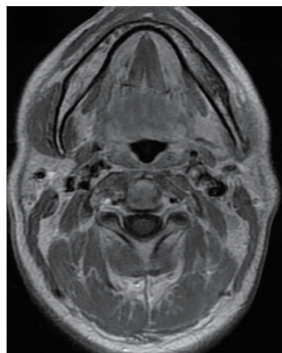
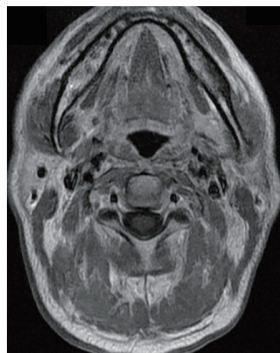


Fig. 11 A failed case of a false positive in a 33-year-old male who presented with recurrence in his left para-nasopharyngeal space involving the left retropharyngeal space, palatine fossa, foramen lacerum, oval foramen and cavernous sinus. The patient was diagnosed with nasopharyngeal carcinoma (T4N0M0 IV stage [AJCC Staging System 8th version]) at the initial presentation. He received radical concurrent chemoradiotherapy. Follow-up H&N MR scans were routinely performed at every visit. This patient did not receive any treatment during a follow-up period of over 2 years. **a)** The patient achieved complete remission of the primary tumor 3 months posttreatment. Moreover, **b)** 6 months, **c)** 12 months after the first course of treatment, the treated survivor remained in complete remission from the disease. **d)** Soft tissue with an enhanced signal was observed in the left cavernous sinus on the follow-up MR image upon 15 months after treatment. As there was no obvious extension, the lesion was not labelled recurrence by the radiologist or clinician at this point. The patient underwent close follow-up at the clinician's recommendation. **e)** Four months after the previous scan, the lesion in the left cavernous sinus progressed with a more enhanced signal but no enlargement in size. Again, the lesion was not labelled as recurrence by the radiologist or clinician at this point. The patient received no further intervention. **f)** From the last scan available for this patient, the 'suspected' lesion progressed, enlarging in size, showing a more enhanced signal, and demonstrating involvement of more adjacent structures. Biopsy was not possible due to the intracranial location of the lesion. Eventually, metabolic functional imaging of this patient confirmed that the hypermetabolic 'lesion' with a standard uptake value of 19.3 in the left cavernous sinus was a recurrence. **g)** Illustration of the risk prediction. In this case, the RAIN model predicted recurrence above the threshold with a lead time of 9 months. This is likely to be a case that would warrant further investigation, such as with metabolic functional imaging. T, tumor; N, lymph node(s); M, metastasis; AJCC, American Joint Committee on Cancer; H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

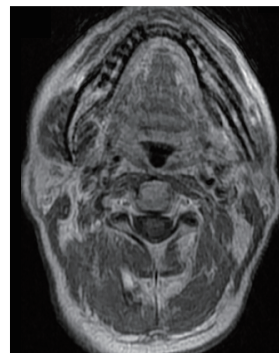
a) 3 months
Complete remission



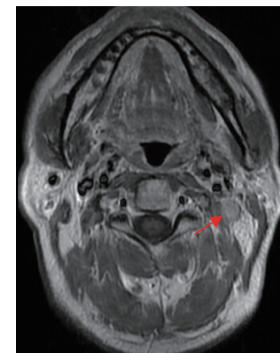
b) 6 months
No recurrence



c) 12 months
No recurrence



d) 27 months
Recurrence



e)

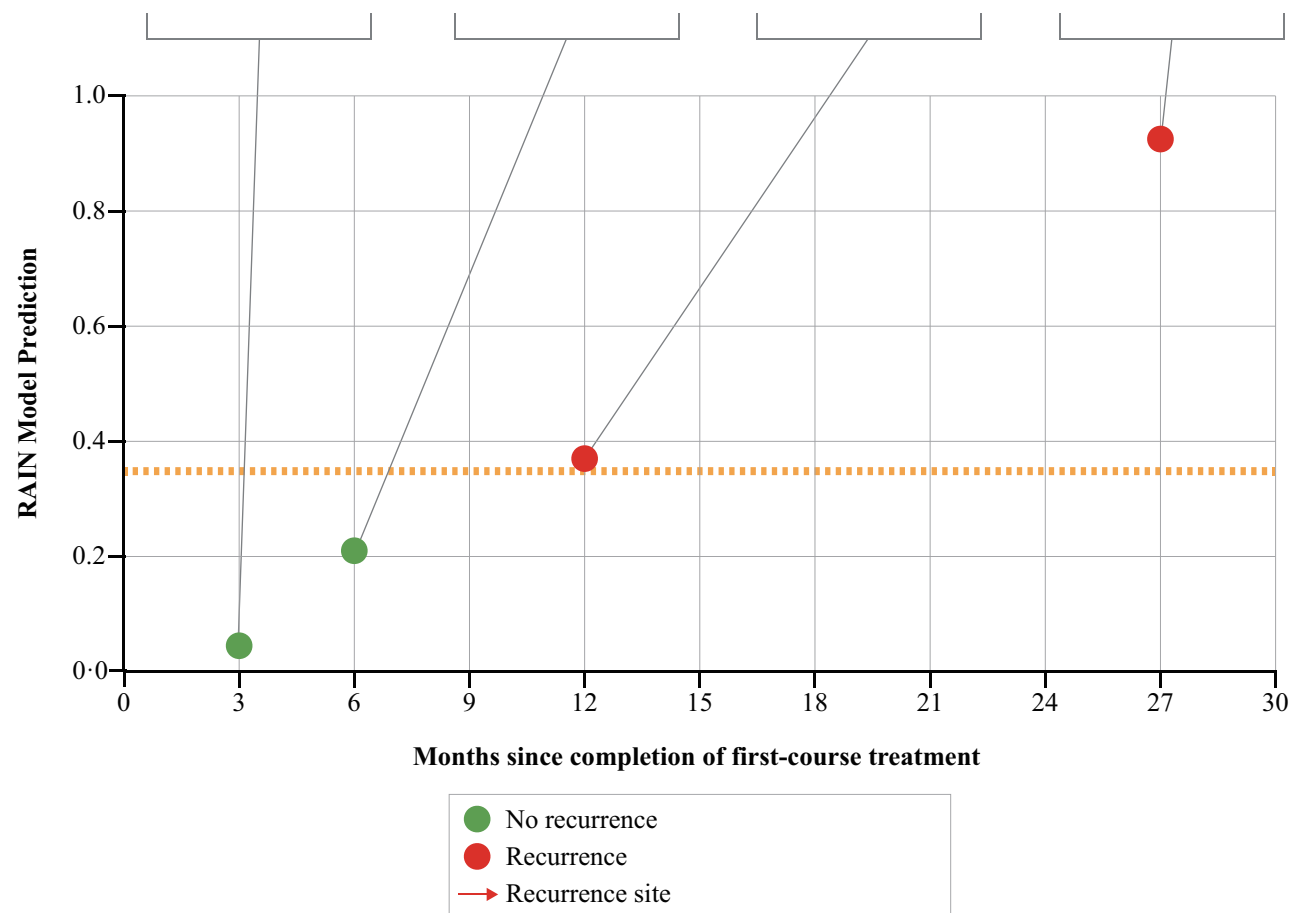
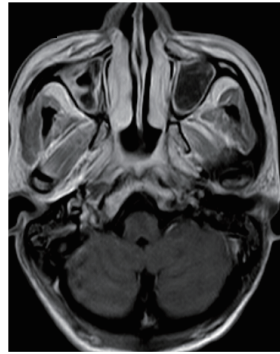
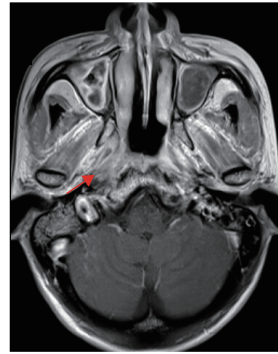


Fig. 12 A failed case of a false positive in a 59-year-old male presented with recurrence in his left cervical lymph node (IIb region). The patient was diagnosed with nasopharyngeal carcinoma (T2N2M0 III stage [AJCC Staging System 8th version]) at the initial presentation. He received radical neoadjuvant chemotherapy and concurrent chemoradiotherapy. Follow-up H&N MR scans were routinely performed at every visit. **a)** The patient achieved complete remission of the primary tumor 3 months posttreatment. **b)** At 6 months and **c)** 12 months after the first course of treatment, the treated survivor remained in complete remission from the disease. **d)** An enlarged lymph node was observed in the left cervical IIb region on follow-up MR image upon 27 months after treatment. The patient underwent selective lymph node dissection in the left neck. The histopathologic examination of the dissected lymph nodes in the left neck showed metastatic properties of nasopharyngeal carcinoma. **e)** Risk prediction graph showing the risk of recurrence rising slightly above the threshold at the third scan prior to progression. Here, the RAIN model predicted recurrence above the threshold with a lead time of 15 months. This is likely to be a case that warranted a closer follow-up regimen. T, tumor; N, lymph node(s); M, metastasis; AJCC, American Joint Committee on Cancer; H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

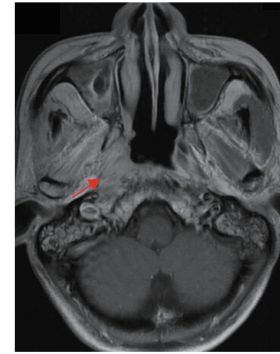
a) 3 months
Complete remission



b) 12 months
Recurrence (rT2N1 rII stage)



c) 23 months
Recurrence (rT4N1 rIV stage)



d)

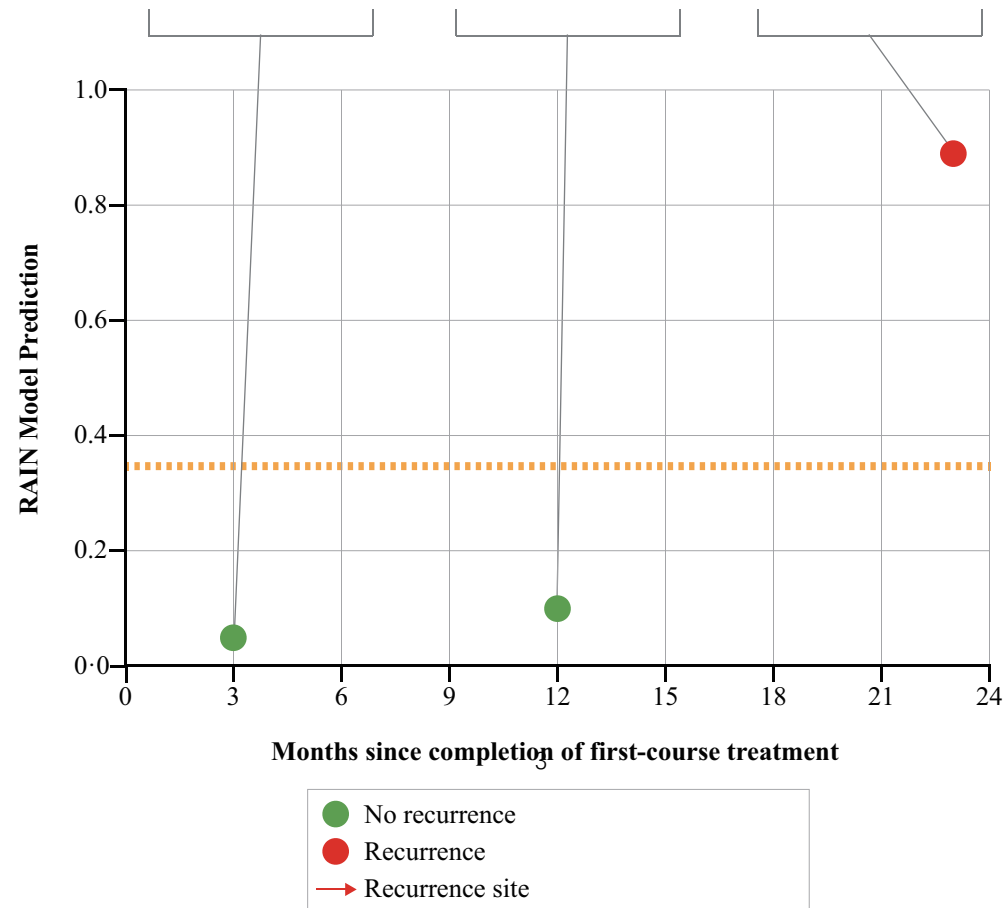
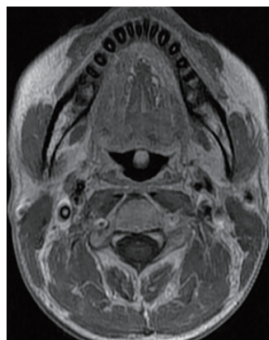
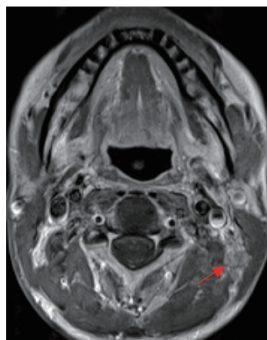


Fig. 13 A failed case of a false negative in a 38-year-old female presented with recurrence in her right parietal wall and right lateral wall of the nasopharynx involving the right musculus levator veli palatini, musculus tensor veli palatini, musculus longus capitis, medial pterygoid, lateral pterygoid muscle, foramen lacerum, oval foramen and cavernous sinus. The patient was diagnosed with nasopharyngeal carcinoma (T3N2M0 III stage [AJCC Staging System 8th version]) at the initial presentation. She received radical concurrent chemoradiotherapy. Follow-up H&N MR scans were routinely performed at every visit. **a)** The patient achieved complete remission of the primary tumor 3 months posttreatment. **b)** Twelve months after the first course of treatment, the treated survivor was misdiagnosed as having ‘no recurrence’. However, the MR scan showed that the lesions in the right lateral wall of the nasopharynx were enlarged with an enhanced signal involving the right musculus levator veli palatini, musculus tensor veli palatini, and inferior nasal concha. However, the patient was misdiagnosed. She received no further intervention and continued routine follow-up. **c)** Eleven months after the previous scan, the ‘missed’ lesions progressed, enlarging in size and involving more adjacent structures. At this point, the RAIN model predicted that this patient has ‘recurrence’. **d)** Risk prediction graph showing an incorrect low risk of recurrence up to the recurrence date. At the second scan, the original radiologist reported that a ‘suspected’ lesion in the right lateral wall of the nasopharynx needed further examination or closer follow-up. T, tumor; N, lymph node(s); M, metastasis; AJCC, American Joint Committee on Cancer; H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

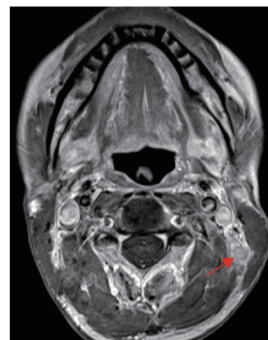
a) 3 months
Complete remission



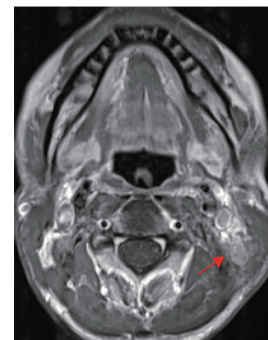
b) 7 months
Recurrence (rT0N1 rII stage)



c) 9 months
Recurrence (rT0N1 rII stage)



d) 13 months
Recurrence (rT0N1 rII stage)



e)

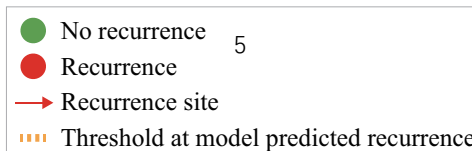
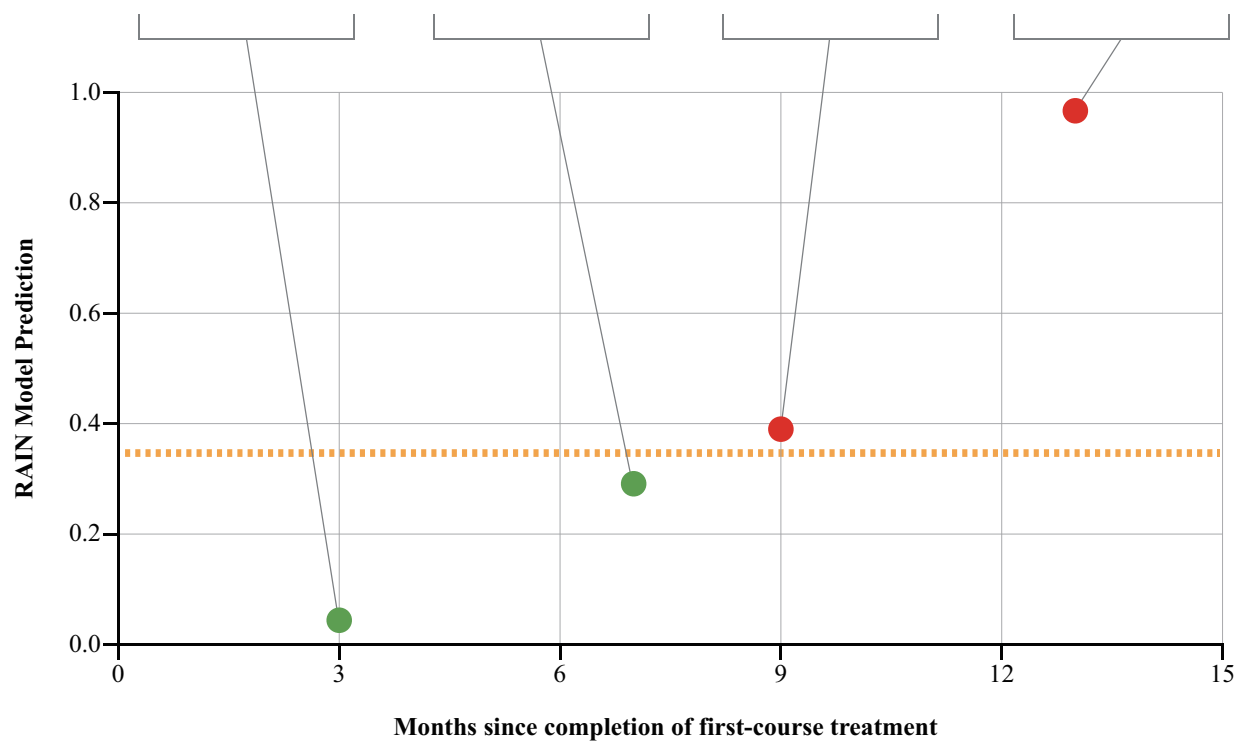


Fig. 14 A failed case of a false negative in a 46-year-old male who presented with recurrence in the left cervical lymph node (IIb region). The patient was diagnosed with nasopharyngeal carcinoma (T3N2M0 III stage [AJCC Staging System 8th version]) at the initial presentation. He received neoadjuvant chemotherapy and concurrent chemoradiotherapy. Follow-up H&N MR scans were routinely performed at every visit. **a)** The patient achieved complete remission of the primary tumor 3 months posttreatment. **b)** Seven months after the first course of treatment, an enlarged lymph node was observed in the left cervical IIb region on the follow-up MR image. However, the patient received anti-inflammatory therapy. **c)** Two months after the previous scan, the ‘missed’ lesions remained stable. Thus, the patient received no further investigation, but a closer follow-up visit was suggested. **d)** Then, 13 months after treatment, the ‘missed’ lymph node progressed, enlarging in size and showing extracapsular spread. The patient underwent radical lymph node dissection in the left neck. The histopathologic examination of the dissected lymph nodes in the left neck showed metastatic properties of nasopharyngeal carcinoma. **e)** Risk prediction graph showing the risk of recurrence rising slightly below the threshold at the second scan. At the second scan, the original radiologist reported that a ‘suspected’ enlarged lymph node in the left IIb cervical region that needed further examination. T, tumor; N, lymph node(s); M, metastasis; AJCC, American Joint Committee on Cancer; H&N, head and neck; MR, magnetic resonance; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

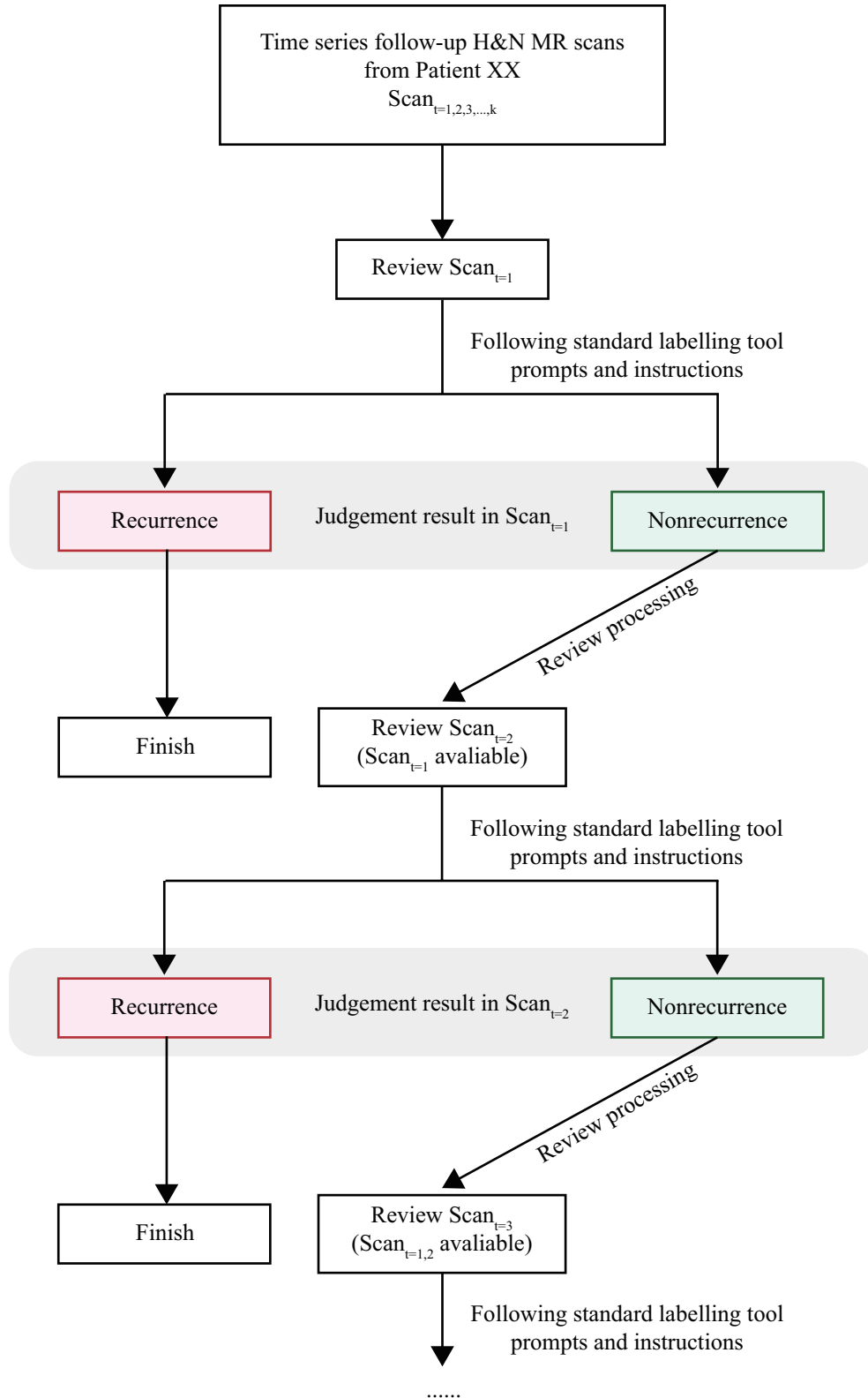


Fig. 15 Scan review process. In this retrospective and prospective study, follow-up H&N MR scans from patients with treated NPC were presented to a rotating panel of readers at the individual level in chronological order. The primary task was to identify recurrence on H&N MR images. In the

retrospective study, a final decision was made in consensus by all of the experts if there was uncertainty.

In the prospective study, the readers independently reviewed the scans. A follow-up H&N MR scan was labelled either positive or negative for recurrence following standard labelling tool prompts and the instructions listed in **Supplementary Table 10**. H&N, head and neck; MR, magnetic resonance; t, time.

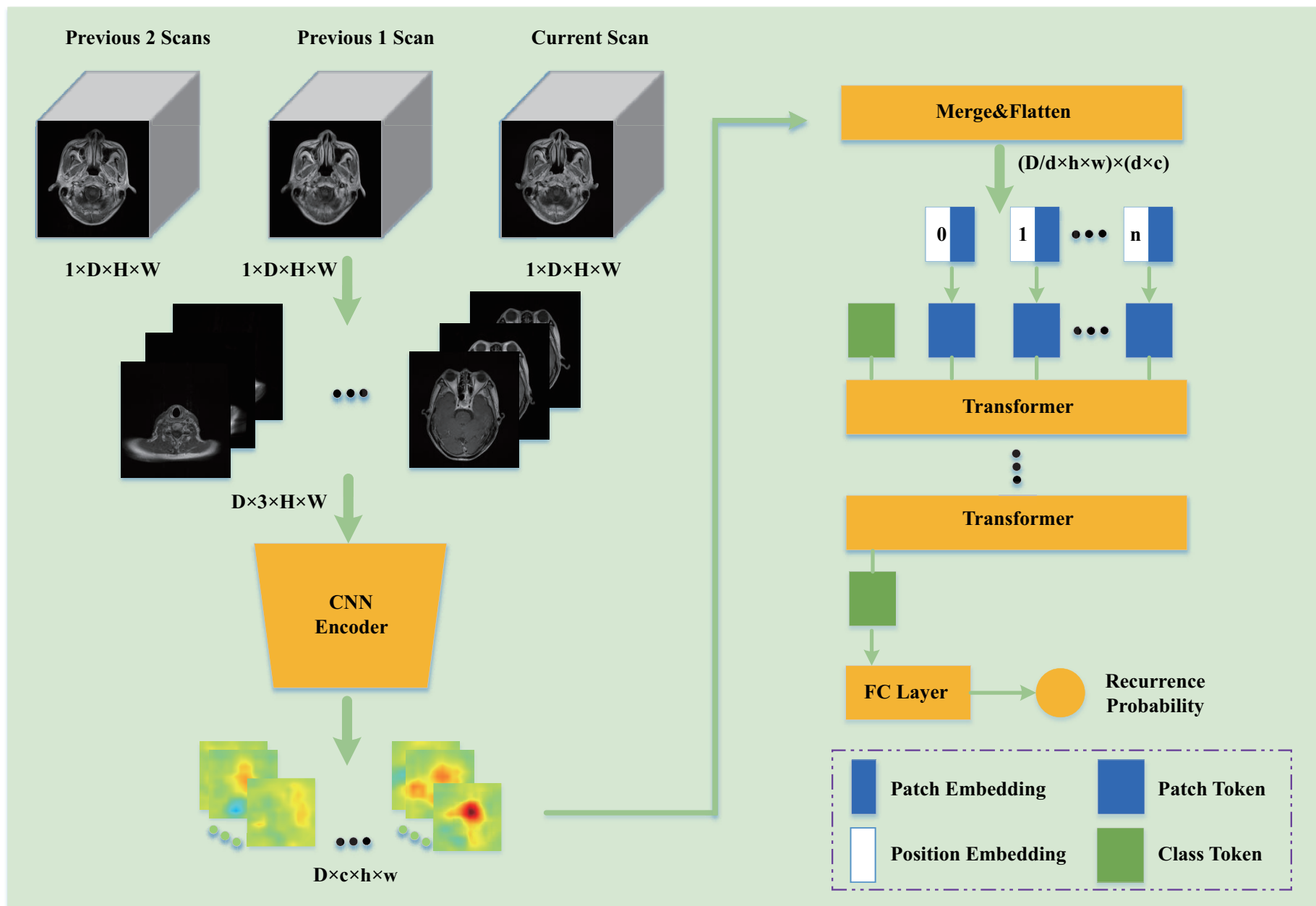


Fig. 16 Flowchart of the RAIN model. CNN, convolutional neural network; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

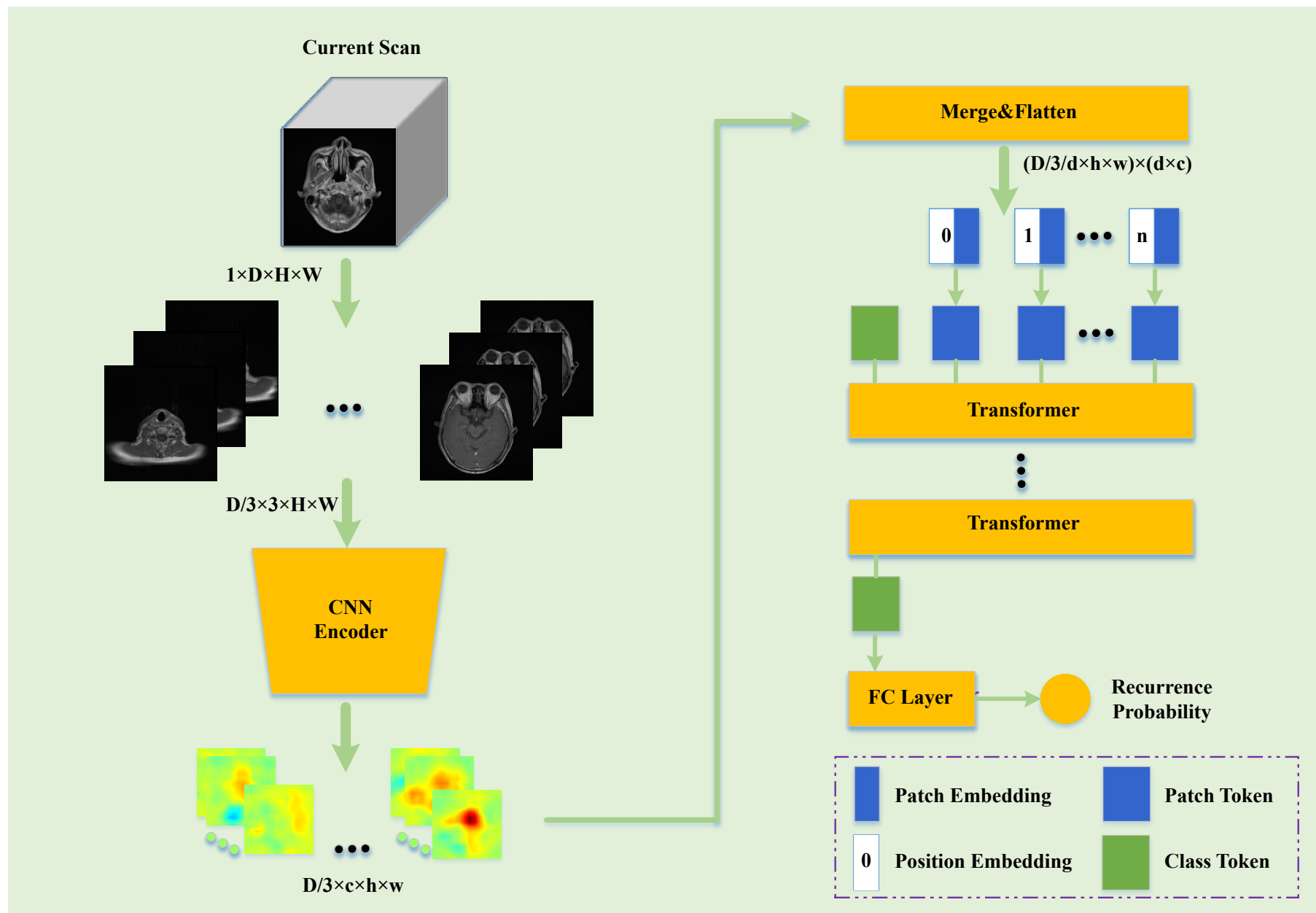


Fig. 17 Flowchart of the single scan model. CNN, convolutional neural network

Part III. Supplementary Tables

Table 1. Imaging characteristics of the confirmed recurrences.

	Training (<i>n</i> = 6 374)	Tuning (<i>n</i> = 815)	Internal Validation (<i>n</i> = 778)	External Validation (<i>n</i> = 469)	Prospective Validation (<i>n</i> = 1 776)
No. of recurrence scans	634 (10.0)	79 (9.7)	79 (10.2)	44 (9.4)	67 (3.8)
Diagnosis of recurrence					
Histopathological	522 (82.3)	58 (73.4)	62 (78.5)	24 (54.5)	56 (83.6)
Functional imaging	45 (7.1)	5 (6.3)	4 (5.1)	1 (2.3)	7 (10.4)
Clinical	67 (10.6)	16 (20.3)	13 (16.4)	19 (43.2)	4 (6.0)
Recurrent tumor type					
Local	327 (51.6)	43 (54.4)	32 (40.5)	30 (68.2)	40 (59.7)
Regional	202 (31.9)	21 (26.6)	26 (32.9)	11 (25.0)	14 (20.9)
Locoregional	105 (16.5)	15 (19.0)	21 (26.6)	3 (6.8)	13 (19.4)
Recurrent T stage*					
rT0	202 (31.9)	21 (26.6)	26 (32.9)	11 (25.0)	14 (20.9)
rT1	80 (12.6)	10 (12.7)	11 (13.9)	5 (11.4)	7 (10.4)
rT2	88 (13.9)	13 (16.5)	12 (15.2)	6 (13.6)	7 (10.4)
rT3	161 (25.4)	25 (31.6)	17 (21.5)	14 (31.8)	25 (37.3)
rT4	103 (16.2)	10 (12.7)	13 (16.5)	8 (18.2)	14 (20.9)
Recurrent N stage*					
rN0	327 (51.6)	43 (54.4)	32 (40.5)	30 (68.2)	40 (59.7)
rN1	243 (38.3)	34 (43.1)	35 (44.3)	7 (15.9)	23 (3.4)
rN2	34 (5.4)	0 (0.0)	5 (6.3)	2 (4.5)	1 (1.5)
rN3	30 (4.7)	2 (2.5)	7 (8.9)	5 (11.4)	3 (4.5)
Recurrent clinical stage*					
rI	68 (10.7)	10 (12.7)	7 (8.9)	4 (9.1)	5 (7.4)
rII	252 (39.7)	32 (40.5)	34 (43.0)	12 (27.3)	20 (29.9)
rIII	183 (28.9)	25 (31.6)	19 (24.1)	16 (36.3)	25 (37.3)
rIV	131 (20.7)	12 (15.2)	19 (24.1)	12 (27.3)	17 (25.4)

Note. —Unless otherwise specified, data are presented as raw data with percentage in the parentheses. rT, recurrent tumor; rN, recurrent metastatic node(s).

*The 8th edition of the American Joint Committee on Cancer (AJCC) guidelines was used for recurrent tumor, metastatic lymph node(s) staging.

Table 2. Imaging characteristics of the earliest visible recurrences.

	Training (<i>n</i> = 6 374)	Tuning (<i>n</i> = 815)	Internal Validation (<i>n</i> = 778)	External Validation (<i>n</i> = 469)	Prospective Validation (<i>n</i> = 1 776)
No. of recurrence scans	634 (10.0)	79 (9.7)	79 (10.2)	44 (9.4)	67 (3.8)
Recurrent tumor type					
Local	350 (55.2)	46 (58.3)	34 (43.0)	31 (70.5)	42 (62.7)
Regional	203 (32.0)	22 (27.8)	27 (34.2)	11 (25.0)	16 (23.9)
Locoregional	81 (12.8)	11 (13.9)	18 (22.8)	2 (4.5)	9 (13.4)
Recurrent T stage*					
rT0	203 (32.0)	22 (27.8)	27 (34.2)	11 (25.0)	16 (23.9)
rT1	106 (16.7)	13 (16.5)	12 (15.2)	6 (13.6)	12 (17.9)
rT2	112 (17.7)	15 (19.0)	13 (16.5)	6 (13.6)	7 (10.4)
rT3	143 (22.6)	23 (29.1)	22 (27.8)	13 (29.6)	24 (35.9)
rT4	70 (11.0)	6 (7.6)	5 (6.3)	8 (18.2)	8 (11.9)
Recurrent N stage*					
rN0	350 (55.2)	46 (58.3)	34 (43.0)	31 (70.5)	42 (62.7)
rN1	236 (37.2)	31 (39.2)	36 (45.6)	8 (18.2)	23 (34.3)
rN2	24 (3.8)	0 (0.0)	3 (3.8)	1 (2.2)	1 (1.5)
rN3	34 (5.4)	2 (2.5)	6 (7.6)	4 (9.1)	1 (1.5)
Recurrent clinical stage*					
rI	91 (14.3)	13 (16.5)	8 (10.1)	6 (13.6)	11 (16.4)
rII	289 (45.6)	35 (44.3)	38 (48.1)	13 (29.6)	22 (32.8)
rIII	161 (25.4)	23 (29.1)	23 (29.1)	14 (31.8)	25 (37.3)
rIV	93 (14.7)	8 (10.1)	10 (12.7)	11 (25.0)	9 (13.4)

Note. —Unless otherwise specified, data are presented as raw data with percentage in the parentheses. rT, recurrent tumor; rN, recurrent metastatic node(s).

*The 8th edition of the American Joint Committee on Cancer (AJCC) guidelines was used for recurrent tumor, metastatic lymph node(s) staging.

Table 3. Performance of the ResNet18 neural network for identifying scans with recurrence in the training and tuning cohorts.

	AUC*	Threshold value [†]	Accuracy (%)	Precision (%)	Recall (%)
Confirmed recurrence scan					
Training Cohort (<i>n</i> = 634/6 374)					
ResNet18 (single scan task)	0.999 (0.998, 1.000)	0.4890	89.9 (5 730/6 374) [89.2, 90.6]	49.6 (634/1 278) [46.9, 52.3]	100.0 (634/634) [100.0, 100.0]
ResNet18 (sequential scan task)	0.995 (0.994, 0.997)	0.5000	94.2 (6 004/6 374) [93.6, 94.8]	63.3 (631/997) [60.4, 66.3]	99.5 (631/634) [99.0, 100.0]
Tuning Cohort (<i>n</i> = 79/815)					
ResNet18 (single scan task)	0.852 (0.827, 0.877)	0.4890	77.7 (633/815) [74.8, 80.5]	27.1 (61/225) [21.3, 32.9]	77.2 (61/79) [68.0, 86.5]
ResNet18 (sequential scan task)	0.982 (0.974, 0.990)	0.5000	91.9 (749/815) [90.0, 93.8]	54.7 (75/137) [46.4, 63.1]	94.9 (75/79) [90.1, 99.8]
Earliest visible recurrence					
Training Cohort (<i>n</i> = 988/6 374)					
ResNet18 (single scan task)	0.999 (0.999, 1.000)	0.5040	92.6 (5 902/6 374) [92.0, 93.2]	67.7 (988/1 459) [65.3, 70.1]	100.0 (988/988) [100.0, 100.0]
ResNet18 (sequential scan task)	0.998 (0.998, 0.999)	0.3450	92.7 (5 909/6 374) [92.1, 93.4]	68.0 (988/1 453) [65.6, 70.4]	100.0 (988/988) [100.0, 100.0]
Tuning Cohort (<i>n</i> = 128/815)					
ResNet18 (single scan task)	0.802 (0.778, 0.827)	0.5040	72.3 (589/815) [69.2, 75.3]	32.7 (93/284) [27.3, 38.2]	72.7 (93/128) [64.9, 80.4]
ResNet18 (sequential scan task)	0.884 (0.866, 0.902)	0.3450	82.2 (670/815) [79.6, 84.8]	46.2 (104/225) [39.7, 52.7]	81.2 (104/128) [74.5, 88.0]

Note. —Unless otherwise specified, data are presented as percentages with numerator/denominator in the parentheses and 95% confidence intervals in brackets. AUC, area under the receiver operating characteristic curve; CI, confidence interval.

*Data are presented as raw data with 95% confidence intervals in parentheses.

[†]Data are numbers. The threshold value was set to dichotomize the magnetic resonance scans into those showing recurrence or nonrecurrence while balancing the tuning cohort by leveraging the Youden method on the receiver operating characteristic curve.

Table 4. Performance and diagnostic accuracy of the RAIN model (sequential scan model) vs single scan model in identifying scans with confirmed recurrence in the retrospective study.

	Single scan model (%)	RAIN (sequential scan model) (%)
Retrospective Study		
Internal Validation Cohort (<i>n</i> = 79/778)		
AUC [†]	0.860 (0.834, 0.886)	0.990 (0.984, 0.996)
Accuracy	75.1 (584/778) [72.0, 78.1]	92.9 (723/778) [91.1, 94.7]
Sensitivity	78.5 (62/79) [69.4, 87.5]	97.5 (77/79) [94.0, 100.0]
Specificity	74.7 (522/699) [71.4, 77.9]	92.4 (646/699) [90.5, 94.4]
PPV	25.9 (62/239) [20.4, 31.5]	59.2 (77/130) [50.8, 67.7]
NPV	96.8 (522/539) [95.4, 98.3]	99.7 (646/648) [99.3, 100.0]
External Validation Cohort (<i>n</i> = 44/469)		
AUC [†]	0.671 (0.616, 0.727)	0.952 (0.939, 0.966)
Accuracy	25.4 (119/469) [21.4, 29.3]	92.5 (434/469) [90.2, 94.8]
Sensitivity	93.2 (41/44) [85.7, 100.0]	76.5 (34/44) [64.8, 88.1]
Specificity	18.3 (78/425) [14.7, 22.0]	94.4 (400/425) [92.2, 96.5]
PPV	10.6 (41/387) [7.5, 13.6]	60.9 (34/55) [49.0, 72.9]
NPV	96.3 (78/81) [92.2, 100.0]	97.2 (400/411) [95.6, 98.8]

Note. —Unless otherwise specified, data are presented as percentages with numerator/denominator in the parentheses and 95% confidence intervals in brackets. AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; CI, confidence interval.

[†]Data are numbers with 95% confidence intervals in parentheses.

The threshold value was set to dichotomize the magnetic resonance scans into those showing recurrence or nonrecurrence while balancing in the tuning cohort by leveraging the Youden method on the receiver operating characteristic curve.

Table 5. Diagnostic accuracy of RAIN Model (sequential scan model) in identifying MR scans with recurrence in retrospective validation cohorts stratified by different clinical factors.

	AUC*	Accuracy (%)	Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)
Retrospective study						
Internal Validation Cohort (<i>n</i> = 118/778)						
Sex						
Female	0.919 (0.865, 0.974)	83.0 (156/188) [77.6, 88.4]	75.0 (18/24) [56.3, 93.7]	84.1 (138/164) [78.5, 89.8]	40.9 (18/44) [25.8, 56.0]	95.8 (138/144) [92.5, 99.1]
Male	0.914 (0.885, 0.944)	81.4 (480/590) [78.2, 84.5]	83.0 (78/94) [75.2, 90.7]	81.0 (402/496) [77.6, 84.5]	45.3 (78/172) [37.8, 52.9]	96.2 (402/418) [94.3, 98.0]
Age, years						
< 45	0.908 (0.865, 0.950)	80.7 (347/430) [77.0, 84.4]	80.0 (44/55) [69.1, 90.9]	80.8 (303/375) [76.8, 84.8]	37.9 (44/116) [29.0, 46.9]	96.5 (303/314) [94.5, 98.5]
≥ 45	0.925 (0.893, 0.956)	83.0 (289/348) [79.1, 87.0]	82.5 (52/63) [72.9, 92.2]	83.2 (237/285) [78.8, 87.5]	52.0 (52/100) [42.0, 62.0]	95.6 (237/248) [93.0, 98.1]
Recurrence time, years [†]						
< 2	0.931 (0.900, 0.963)	82.2 (600/730) [79.4, 85.0]	85.7 (60/70) [77.3, 94.1]	81.8 (540/660) [78.9, 84.8]	33.3 (60/180) [26.4, 40.3]	98.2 (540/550) [97.1, 99.3]
3–5	0.893 (0.854, 0.933)	81.4 (576/708) [78.5, 84.2]	75.0 (36/48) [62.3, 87.7]	81.8 (540/660) [78.9, 84.8]	23.1 (36/156) [16.4, 29.8]	97.8 (540/552) [96.6, 99.0]
Recurrent tumor type						
Local	0.926 (0.894, 0.958)	81.9 (585/714) [79.1, 84.8]	83.3 (45/54) [73.1, 93.6]	81.8 (540/660) [78.9, 84.8]	27.3 (45/165) [20.4, 34.1]	98.4 (540/549) [97.3, 99.4]
Regional	0.901 (0.856, 0.945)	81.6 (571/700) [78.7, 84.5]	77.5 (31/40) [64.0, 91.0]	81.8 (540/660) [78.9, 84.8]	20.5 (31/151) [14.0, 27.0]	98.4 (540/549) [97.3, 99.4]
Locoregional	0.918 (0.854, 0.982)	81.9 (560/684) [79.0, 84.8]	83.3 (20/24) [67.3, 99.4]	81.8 (540/660) [78.9, 84.8]	14.3 (20/140) [8.4, 20.2]	99.3 (540/544) [98.5, 100.0]
Recurrent clinical stage [§]						
rI–II	0.893 (0.856, 0.930)	81.2 (592/729) [78.4, 84.0]	75.4 (52/69) [64.9, 85.8]	81.8 (540/660) [78.9, 84.8]	30.2 (52/172) [23.3, 37.2]	97.0 (540/557) [95.5, 98.4]
rIII–IV	0.948 (0.919, 0.976)	82.4 (584/709) [79.6, 85.2]	89.8 (44/49) [81.0, 98.6]	81.8 (540/660) [78.9, 84.8]	26.8 (44/164) [20.0, 33.7]	99.1 (540/545) [98.3, 99.9]
External Validation Cohort (<i>n</i> = 56/469)						
Sex						
Female	0.910 (0.829, 0.991)	86.8 (145/167) [81.6, 92.0]	75.0 (12/16) [51.2, 98.8]	88.1 (133/151) [82.9, 93.3]	40.0 (12/30) [21.4, 58.6]	97.1 (133/137) [94.2, 99.9]
Male	0.894 (0.840, 0.948)	83.8 (253/302) [79.6, 88.0]	77.5 (31/40) [64.0, 91.0]	84.7 (222/262) [80.3, 89.1]	43.7 (31/71) [31.8, 55.5]	96.1 (222/231) [93.6, 98.6]
Age, years						
< 45	0.839 (0.747, 0.932)	86.3 (220/255) [82.0, 90.5]	63.6 (14/22) [41.8, 85.5]	88.4 (206/233) [84.3, 92.6]	34.1 (14/41) [19.0, 49.3]	96.3 (206/214) [93.7, 98.8]

≥ 45	0.939 (0.904, 0.974)	83.2 (178/214) [78.1, 88.2]	85.3 (29/34) [72.8, 97.8]	82.8 (149/180) [77.2, 88.3]	48.3 (29/60) [35.3, 61.4]	96.7 (149/154) [93.9, 99.6]
Recurrence time, years [†]						
< 2	0.885 (0.829, 0.941)	84.8 (385/454) [81.5, 88.1]	73.2 (30/41) [59.0, 87.3]	86.0 (355/413) [82.6, 89.3]	34.1 (30/88) [24.0, 44.2]	97.0 (355/366) [95.2, 98.8]
3–5	0.942 (0.894, 0.990)	86.0 (368/428) [82.7, 89.3]	86.7 (13/15) [67.2, 100.0]	86.0 (355/413) [82.6, 89.3]	18.3 (13/71) [9.1, 27.5]	99.4 (355/357) [98.7, 100.0]
Recurrent tumor type						
Local	0.909 (0.857, 0.961)	85.6 (387/452) [82.4, 88.9]	82.0 (32/39) [69.4, 94.7]	86.0 (355/413) [82.6, 89.3]	35.6 (32/90) [25.5, 45.6]	98.1 (355/362) [96.6, 99.5]
Regional	0.868 (0.783, 0.954)	85.0 (364/428) [81.7, 88.4]	60.0 (9/15) [31.9, 88.1]	86.0 (355/413) [82.6, 89.3]	13.4 (9/67) [5.1, 21.8]	98.3 (355/361) [97.0, 99.7]
Locoregional	0.971 (0.939, 1.000)	86.0 (357/415) [82.7, 89.4]	100.0 (2/2) [100.0, 100.0]	86.0 (355/413) [82.6, 89.3]	3.3 (2/60) [0.0, 8.0]	100.0 (355/355) [100.0, 100.0]
Recurrent clinical stage [§]						
rI–II	0.845 (0.765, 0.924)	84.7 (372/439) [81.4, 88.1]	65.4 (17/26) [45.8, 85.0]	86.0 (355/413) [82.6, 89.3]	22.7 (17/75) [13.0, 32.4]	97.5 (355/364) [95.9, 99.1]
rIII–IV	0.948 (0.915, 0.981)	86.0 (381/443) [82.8, 89.2]	86.7 (26/30) [73.8, 99.6]	86.0 (355/413) [82.6, 89.3]	30.9 (26/84) [20.9, 41.0]	98.9 (355/359) [97.8, 100.0]

Note. —Unless otherwise specified, data are presented as percentages with numerator/denominator in the parentheses and 95% confidence intervals in brackets. RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; MR, magnetic resonance; AUC, area under the receiver operating characteristic curve; PPV, positive predictive value; NPV, negative predictive value; r, recurrent.

*Data are numbers with 95% CI in parentheses.

[†]Recurrence time was calculated from the time between the end of the first radiotherapy course and the diagnosis of recurrence.

[§]The 8th edition of the American Joint Committee on Cancer (AJCC) guidelines was used for recurrent tumor, metastatic lymph node(s) staging.

Table 6. Improvement of RAIN model (sequential scan model) on timely diagnosis of recurrence at individual level in the retrospective study.

	Leading time, mean ($\pm$ SD) months *	Increased rate of timely diagnosis of recurrence, <i>n</i> (%)	<i>P</i>	Increased rate of recurrence at early stage (rI–II), <i>n</i> (%) †	<i>P</i>
Training cohort (<i>n</i> = 634)	3.8 ($\pm$ 5.7)	282 (44.5%)	< 0.001	60 (9.5%)	0.003
Tuning cohort (<i>n</i> = 79)	2.8 ($\pm$ 5.0)	39 (32.9%)	< 0.001	4 (5.0%)	0.522
Internal validation cohort (<i>n</i> = 79)	1.5 ($\pm$ 4.4)	27 (17.7%)	0.013	4 (5.1%)	0.523
External validation cohort (<i>n</i> = 44)	1.0 ($\pm$ 3.2)	4 (11.3%)	0.156	2 (4.5%)	0.661

Note. —Unless otherwise specified, data are presented as percentages with numerator/denominator in the parentheses. RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; SD, standard deviation; r, recurrent.

*Data are mean value with $\pm$ SD in the parentheses.

†The 8th edition of the American Joint Committee on Cancer (AJCC) guidelines was used for recurrent tumour, metastatic lymph node(s) staging.

Table 7. False positive alerts.

	Unique rNPC with at least one FP*	Inspection(s) prior to recurrence (%)			Months prior to recurrence (%)		
		1 inspection prior to recurrence	2 inspections prior to recurrence	≥3 inspections prior to recurrence	0–6 months prior to recurrence	6–12 months prior to recurrence	≥12 months prior to recurrence
Number of cases	13	8 (61.5)	2 (15.4)	3 (23.1)	3 (23.1)	3 (23.1)	7 (53.8)

Note. —Unless otherwise specified, data are presented as raw data with percentage in the parentheses. rNPC, recurrent nasopharyngeal carcinoma.

*Data are raw data.

Table 8. Improvement of RAIN model (sequential scan model) and readers on timely diagnosis of recurrence at individual level in the prospective study.

Prospective validation cohort (<i>n</i> = 67)	Leading time, mean (±SD) months *	Increased rate of timely diagnosis of recurrence, <i>n</i> (%)	<i>P</i>	Increased rate of recurrence at early stage (rI–II), <i>n</i> (%) †	<i>P</i>
RAIN Model	3.8 (±5.7)	29 (43.3%)	< 0.001	5 (7.5%)	0.380
Expert Reader	2.5 (±4.6)	24 (35.8%)	< 0.001	3 (4.5%)	0.596
Competent Reader	2.0 (±4.3)	18 (26.8%)	0.002	2 (3.0%)	0.723
Expert Reader	1.8 (±4.2)	14 (20.9%)	0.015	1 (1.5%)	0.859

Note. —Unless otherwise specified, data are presented as percentages with numerator/denominator in the parentheses. RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; SD, standard deviation; r, recurrent.

*Data are mean value with ±SD in the parentheses.

†The 8th edition of the American Joint Committee on Cancer (AJCC) guidelines was used for recurrent tumour, metastatic lymph node(s) staging.

Table 9. The acquisition parameters of H&N MR scanning.

	n (%)	Magnetic field strength (T)	T1-WI					T2-WI				
			TR (ms)	TE (ms)	Pixel size (mm)	FOV (mm)	Slice thickness (mm)	TR (ms)	TE (ms)	Pixel size (mm)	FOV (mm)	Slice thickness (mm)
Hospital 1 (n = 9 743)												
Magnetom Aera; SIEMENS, Germany	1 912 (19.6)	1.5	450	7.6	0.80×0.80	240×225	5.0	7000	77.0	0.60×0.60	240×225	5.0
Panorama HFO; PHILIPS, Netherlands	1 713 (17.6)	1.0	400	8.0	0.80×0.62	240×221	4.0	2500	80.0	0.80×0.62	240×221	4.0
Achieva 3.0T; PHILIPS, Netherlands	1 663 (17.1)	3.0	610	8.0	0.65×0.81	240×240	5.0	2650	89.0	0.65×0.81	240×240	5.0
Discovery MR750w; GE, USA	1 625 (16.7)	3.0	610	8.7	0.60×0.70	240×240	5.0	5836	120.0	0.60×0.70	240×240	5.0
Magnetom Trio Tim; SIEMENS, Germany	1 469 (15.1)	3.0	600	8.8	0.80×0.60	240×240	5.0	5000	89.0	0.80×0.60	240×240	5.0
Ingenia CX; PHILIPS, Netherlands	582 (6.0)	3.0	589	13.0	0.70×0.80	240×265	5.0	2552	100.0	0.75×0.85	240×264	5.0
uMR 780; UNITED IMAGING, China	397 (4.1)	3.0	676	8.3	0.89×0.63	240×240	5.0	4449	113.4	0.89×0.63	240×240	5.0
Pioneer; GE, USA	241 (2.5)	3.0	610	12.6	0.60×0.80	240×240	5.0	9121	85.0	0.60×0.80	240×240	5.0
Premier; GE, USA	47 (0.5)	3.0	781	6.2	0.60×0.80	240×240	5.0	5269	68.0	0.60×0.80	240×240	5.0
uMR 790; UNITED IMAGING, China	46 (0.5)	3.0	636	7.8	0.78×0.63	240×240	5.0	4900	120.9	0.89×0.63	240×240	5.0
Magnetom Espree; SIEMENS, Germany	24 (0.3)	1.5	550	9.7	1.10×0.80	240×216	5.0	3800	87.0	1.10×0.80	240×216	5.0
uMR 560; UNITED IMAGING, China	13 (0.1)	1.5	642	11.1	0.89×0.63	240×231	5.0	3700	95.4	0.89×0.63	240×234	5.0
Magnetom Prisma; SIEMENS, Germany	6 (0.1)	3.0	540	13.0	0.60×0.60	240×240	5.0	4430	78.0	0.60×0.60	240×240	5.0
Signa Architect; GE, USA	5 (0.1)	3.0	559	7.5	0.70×0.80	240×216	5.0	6334	85.0	0.60×0.80	240×216	5.0
Hospital 2 (n = 469)												
Achieva; PHILIPS, Netherlands	469 (100.0)	1.5	545	14.0	0.70×1.05	230×230	4.0	3193	80	1.01×1.24	230×230	5.0

Note. —Unless otherwise specified, data are presented as n (%) or metric value. H&N, head and neck; MR, magnetic resonance; T1-WI, T1-weighted imaging; T2-WI, T2-weighted imaging; TR, repetition time; TE, echo time; FOV, field of view.

Table 10. Labelling tool prompts and the instructions.

Patient ID: XX Age: XX Sex: XX DOB: XX Date of treatment completion: dd/mm/yyyy	
Questions	Possible answers (mark the best item with tick in questions with multiple-choices)
Has this patient achieved complete remission (CR)* from (chemo)radiotherapy?	<u>Yes</u> : if there is no visible tumor in the nasopharynx (including the adjacent tissue and structure), bilateral retropharyngeal lymph node regions and bilateral cervical lymph node regions <u>Possibly</u> [†] : if there is suspected residual tumor in the nasopharynx (including the adjacent tissue and structure), bilateral retropharyngeal lymph node regions and bilateral cervical lymph node regions <u>No</u> [†] : if there is apparent residual tumor in the nasopharynx (including the adjacent tissue and structure), bilateral retropharyngeal lymph node regions and bilateral cervical lymph node regions
Does an event of recurrence occur in this patient within the follow-up period?	<u>Yes</u> : if there is any biopsy-, functional imaging-, or radiological evidence confirming recurrence in a local, regional or locoregional site <u>No</u> [†] : if there is no evidence of recurrence
Does an event of recurrence occur in this patient within available follow-up period?	<u>Yes</u> : if there is any biopsy-, functional imaging-, or radiological evidence confirming recurrence in a local, regional or locoregional site <u>No</u> [†] : if there is no evidence of recurrence
Can you specify the site(s) of recurrence on the MR images?	<u>Local</u> : visible recurrent tumor in the nasopharynx (including the adjacent tissue and structure) <u>Regional</u> : visible of recurrent tumor in the lateral or bilateral retropharyngeal lymph node region(s) <u>Locoregional</u> : visible of recurrent tumor in the lateral or bilateral cervical lymph node region(s)
Please designate the date of the MR scan with the earliest visible recurrence:	
Confidence of designation of the scan with the earliest visible recurrence:	<u>5</u> : most certain about the condition <u>4</u> : <u>3</u> : <u>2</u> : <u>1</u> : least certain about the condition

The information on the patient listed in the first row, including anonymized identity, age, sex, date of birth, and date of completion of the first-course treatment, was available to the reader. The readers were required to complete the questionnaire for each patient. ID, identity; DOB, date of birth; CR, complete remission; MR, magnetic resonance.

*The Response Evaluation Criteria in Solid Tumors (RECIST; version 1.1) is used in the radiological efficacy evaluation.

[†]If these answers were selected, the remaining questions can be skipped.

Table 11. Architecture of the model.

Layers	Output Size	RAIN Model	
res1	224×224	7×7 conv, stride 2	
res2	112×112	3×3 max pool, stride 2	
		3×3 conv 3×3 conv	$\times 2$
res3	56×56	3×3 conv 3×3 conv	$\times 2$
res4	28×28	3×3 conv 3×3 conv	$\times 2$
res5	14×14	3×3 conv 3×3 conv	$\times 2$
res6	14×14	1×1 conv	
Merge & Flatten	2940	-	
Transformer	2941	hidden size	256
		mlp size	1024
		layers	4
		heads	6
Classification Layer	1	1D fully-connected, sigmoid	

Part IV. Study Protocol

Title: Application of deep learning in precise detection of recurrence in nasopharyngeal carcinoma: a multicenter observational study

Chinese Clinical Trial Registry Identifier: ChiCTR2200056595

Date: 2021.01.01

Table of Contents

1.0 TRIAL SUMMARY	
2.0 TRIAL DESIGN	
2.1 Trial Design	
2.2 Trial Diagram	
3.0 OBJECTIVE(S) & HYPOTHESIS(ES)	
3.1 Primary Objective(s) & Hypothesis(es)	
3.2 Secondary Objective(s) & Hypothesis(es)	
4.0 BACKGROUND & RATIONALE	
4.1 Background	
4.2 Rationale	
4.2.1 Rationale for the Trial	
4.2.2 Rationale for Endpoints	
5.0 METHODOLOGY	
5.1 Entry Criteria	
5.1.1 Subject Inclusion Criteria	
5.1.2 Subject Exclusion Criteria	
5.1.3 Subjective Withdrawal/Discontinuation Criteria	
5.2 Baseline Characteristics Collection and Serial Number Generation	
5.3 Imaging Data Collection and Process	
5.3.1 Magnetic Resonance Imaging Collection and Anonymization	
5.3.2 Magnetic Resonance Imaging Quality Control	
5.3.3 Magnetic Resonance Imaging Labelling	
5.4 Recurrence Detection by Model	
5.5 Recurrence Detection by Manual Reading	
5.6 Actual Endpoints	
6.0 TRIAL PROCEDURES	
7.0 STATISTICAL ANALYSIS	
7.1 Statistical Analysis Plan Summary	
7.2 Hypotheses/Estimation	
7.3 Analysis Population	
7.4 Statistical Methods for Study Endpoints	
7.4.1 Primary Accuracy Endpoints	
7.4.1.1 Area under curve (AUC)	
7.4.2 Secondary Accuracy Endpoints	
7.4.2.1 Sensitivity	
7.4.2.2 Specificity	
7.4.2.3 Positive Prediction Value	
7.4.2.4 Negative Prediction Value	
7.4.3 Secondary Comparative Endpoints	
7.4.3.1 Delong's test	
7.4.3.2 Student's t test	
7.4.3.3 McNemar test	
7.5 Statistical Methods for Baseline Characteristics and Demographics	

7.6	Sample Size and Power Calculations
8.0	CONFIDENTIALITY AND DATA SHARING PLAN
8.1	Confidentiality of Data
8.2	Data Sharing Plan
9.0	SPONSORS AND COLLABORATORS
9.1	Sponsors and Collaborators
9.2	Responsible Party/Investigator
9.3	Role of Funding
9.4	Ethics Committee
10.0	REFERENCES

1.0 TRIAL SUMMARY

Brief Title	Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma (RAIN)
Official Title	Application of deep learning for the precise detection of recurrence in nasopharyngeal carcinoma: a multicenter observational study
Trial Type	Observational
Time Perspective	Retrospective + Prospective
Medical Context	Diagnosis
Study Population	Patients with complete remission of nonmetastatic nasopharyngeal carcinoma after curative (chemo)radiotherapy
Collaborative Centers	Two centers ● Hospital 1 ● Hospital 2
Study Procedure	Eligible patients between January 2008 and December 2020 will be retrospectively enrolled, and their head and neck magnetic resonance images and clinical information during the follow-up period will be collected. The retrospectively collected cases from January 2021 and June 2021 will be used to develop and validate a deep learning-enabled model to detect earliest visible recurrence on magnetic resonance images. During January 2021 and June 2021, eligible patients will be prospectively enrolled, and their head and neck magnetic resonance images and clinical information will be collected for prospective validation of the model.
Study Groups	Patients will be recruited and divided into retrospective (internal and external) and prospective (internal) validation groups according to the set timepoint (January 2021). Follow-up magnetic resonance images from eligible patients will undergo recurrence detection by the model (Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma, RAIN), which has been previously constructed by investigators.
Sample Size	Approximately 1745 patients will be enrolled.
Estimated Duration	Approximately 12 months from the time the first subject is enrolled until the last subject completes treatment.
Outcome Measures	Primary endpoints: area under the curve (AUC) Secondary endpoints: accuracy, sensitivity, specificity, positive prediction value (PPV), negative prediction value (NPV)

2.0 TRIAL DESIGN

2.1 Trial Design

This is a multicenter observational diagnostic study for the development and retrospective and prospective validation of a deep learning (DL)-based recurrence detection model for diagnosing the earliest radiologically visible recurrence on follow-up magnetic resonance (MR) images of patients with treated nonmetastatic nasopharyngeal carcinoma (NPC).

Approximately 1 500 patients will be retrospectively enrolled from two centers into development, internal and external validation cohorts. Approximately 245 patients will be prospectively enrolled in Hospital 1 as a prospective validation cohort. All patients with nonmetastatic NPC should have received curative (chemo)radiotherapy. Since MR imaging is the standard recommended imaging modality for treatment evaluation, the head and neck (H&N) MR images acquired within 6 months will serve as the standard control for each individual. All patients underwent regular H&N MR imaging examinations during the follow-up period until recurrence or remained in complete remission (CR) for 5 years. Qualified images of H&N MR sequences and other clinical information (e.g., medical history, quantification of plasma Epstein–Barr virus [EBV] deoxyribonucleic acid [DNA], cytology and/or histological biopsy results, and functional imaging [e.g., 18F-Fluorodeoxyglucose positron emission tomography and computed tomography {18F-FDG PET/CT}]) will be collected and stored to a designated cloud platform. All the follow-up H&N MR images available during the follow-up period will be evaluated by a rotating panel of four experts (two radiologists and two oncologists) specialized in NPC from the Hospital 1. After their scans are labelled, patients will be classified as having either rNPC (biopsy-, functional imaging-, or radiologically confirmed recurrence in local, regional or locoregional sites) or non-rNPC proven through follow-up (remained in CR for at least 5 years). Specifically, the follow-up scans prior to labelled scans will be traced backtracks until the MR scan with the earliest radiologically visible recurrence is identified by experts based on the typical and apparent imaging manifestations that could diagnose recurrence. Experts can consider all known clinical information and will make a final decision in consensus. Correspondingly, a follow-up H&N MR scan will be labelled as nonrecurrence, earliest radiologically visible recurrence, or clinical diagnosis of recurrence.

2.2 Trial Diagram

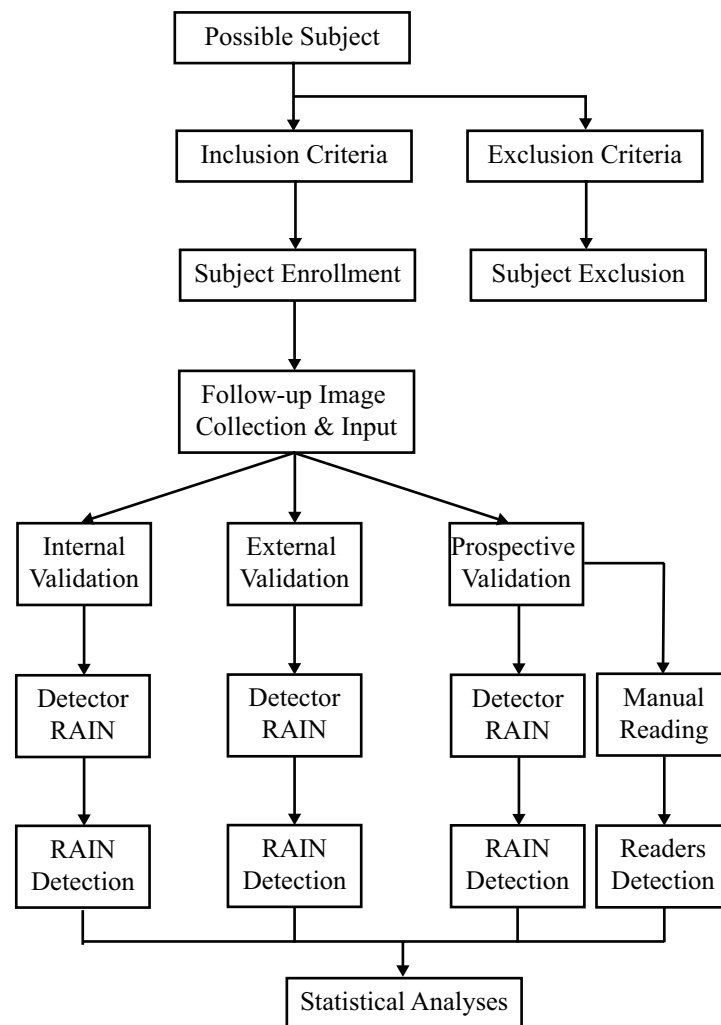


Figure 1. Trial Diagram

3.0 OBJECTIVE(S) & HYPOTHESIS(ES)

3.1 Primary Objective(s) & Hypothesis(es)

(1) **Objective 1:** Evaluate the area under the curve (AUC) of RAIN in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in retrospective validation cohorts.

Hypothesis (H1): RAIN achieves an AUC over 0.85 in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in retrospective validation cohorts.

3.2 Secondary Objective(s) & Hypothesis(es)

(1) **Objective 2:** Compare the AUC of RAIN with that of three readers who manually assess each image in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in a prospective validation cohort.

Hypothesis (H2): RAIN is not inferior to readers in terms of AUC in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in retrospective validation cohorts.

(2) **Objective 3:** Evaluate the accuracy of RAIN in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in retrospective validation cohorts.

(4) **Objective 3:** Compare the accuracy of RAIN with that of three readers who manually assess each image in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in a prospective validation cohort.

(5) **Objective 4:** Evaluate the sensitivity of RAIN in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in retrospective validation cohorts.

(6) **Objective 5:** To compare the sensitivity of RAIN with that of three readers who manually assess each image in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in a prospective validation cohort.

(7) **Objective 6:** Evaluate the specificity of RAIN in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC.

(8) **Objective 7:** Compare the specificity of RAIN with that of three readers who manually assess each image in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in a prospective validation cohort.

(9) **Objective 8:** Evaluate the positive prediction value (PPV) of RAIN in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC.

(10) **Objective 9:** Compare the positive prediction value (PPV) of RAIN with that of three readers who manually assess each image in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC in a prospective validation cohort.

(11) **Objective 10:** Evaluate the negative prediction value (NPV) of RAIN in detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC.

(12) **Objective 11:** Compare the negative prediction value (NPV) of RAIN with that of three readers who manually assess each image in detecting the earliest radiologically visible recurrence on follow-up H&N MR images among patients with treated NPC in a prospective validation cohort.

4.0 BACKGROUND & RATIONALE

4.1 Background

The potential for recurrence makes long-term surveillance vital in managing cancer patients^{1,2}. Nasopharyngeal carcinoma (NPC), which prevails across East and Southeast Asia (3.0 per 100 000)³, is a cancer that requires regular follow-up surveys. Although (chemo)radiotherapy has prolonged survival in patients with NPC (5-year overall survival [OS] rate 87.4%), loco(regional) recurrence (10.0–15.0%) is still a threat among patients with treated NPC, with a 5-year OS rate as low as 26.0% among those with recurrence^{4–6}. This is mainly because recurrence is often advanced with limited treatment options by the time it is detected. Because recurrences in the early stage are often more amenable to local treatment and more curable, preemptive identification might enable timely intervention and drastically ameliorate patient outcomes^{2,7–9}.

Annual head and neck (H&N) magnetic resonance (MR) imaging has been a dominant monitoring scheme in patients with treated NPC^{10,11}. Currently, both the detection and assessment of radiographically visible lesions rely upon visual evaluations. The histological and structural changes induced by radiation, high-level expertise required, changes in subjective judgement, and heavy workload may mean that clinicians inevitably neglect some occult, atypical, and early signs of recurrence or mistake such lesions for benign ‘mimickers’^{12,13}. Thus, some patients might end up with recurrence that, in hindsight, could have been detected on previous scans¹⁴. Hence, improving the precision of recurrence detection in vast set of surveillance images is imperative and may minimize the high costs of missed diagnosis in false negatives and unnecessary procedures resulting in false positives^{12,13}. Although previous models have stratified subgroups with high risk based on assessments of a static tumor status, there is no compelling tool realizing real-time precise recurrence detection^{15–18}. Therefore, a convenient, time- and labour-saving approach is needed. Furthermore, more accurate assessments could potentially lead to future referrals and an earlier diagnosis of recurrence, thereby improving patient outcomes.

Deep learning (DL) exhibits formidable strengths in automating complex image analysis, subtle imaging feature detection and methodology unification for image evaluation through pattern recognition¹⁹. At the dawn of this inspiring technological transformation, an accumulation of advancements in DL engineered for cancer diagnosis have repeatedly been shown to be on-par with and even surpass humans in performance in high-profile studies^{20–22}. Moreover, DL may quantify radiomics information that is undetectable by the naked eye and thereby complement clinical decision-making²². DL possesses the potential to reduce observational oversights and missed errors in detecting recurrence in the surveillance of cancer patients²³. However, there are currently no medical studies exploiting this potential.

4.2 Rationale

4.2.1 Rationale for the Trial

We searched PubMed from database inception to January 1, 2021, for publications on application of artificial intelligence on recurrence surveillance in treated nasopharyngeal carcinoma (NPC). We used the search terms “artificial intelligence” or “deep learning” or “machine learning” and “recurrence”, “surveillance”, “monitoring”, and “nasopharyngeal carcinoma”, without language restrictions. Our search yielded 2 original studies with limited cases: 1 study combined radiomics with a machine learning method to predict in-field recurrences of NPC; and 1 study used a machine learning method with posttreatment positron emission tomography/computed tomography (PET/CT)-based radiomics to differentiate between recurrence and inflammation. We did not identify any studies that investigated the application of artificial intelligence for recurrence surveillance in NPC.

The study is conducted to develop, retrospectively and prospectively evaluate the clinical utility and

generalizability of RAIN for detecting the earliest radiologically visible recurrence on follow-up H&N MR images of patients with treated NPC. The diagnostic performance of RAIN will be evaluated in a multicenter retrospective dataset, and compared to that of three readers who manually assess each image in the prospective dataset, which might potentially provide important evidence for the feasibility and clinical value of integration for the computer-aided detecting tool in healthcare.

4.2.2 Rationale for Endpoints

The primary endpoint in the study is the AUC, a significant indicator of the classification performance of a binary classifier that has been widely used to evaluate model performance in the field of DL.

5.0 METHODOLOGY

5.1 Entry Criteria

5.1.1 Subject Inclusion Criteria

To be eligible for participation in this trial, the subject must:

1. be aged ≥ 18 years with pathologically confirmed nasopharyngeal carcinoma (NPC) at the initial diagnosis.
2. be restaged to I-IV_a (T₁₋₄N₀₋₃M₀) according to the American Joint Committee on Cancer (AJCC) 8th staging classification system ²⁴.
3. have undergone radical radiotherapy (RT) $\pm$ neoadjuvant or concurrent chemotherapy.
4. have achieved radiographically complete remission (CR) of the gross tumor at the primary site and metastatic lymph nodes at the cervical region within 6 months post RT ²⁵.
5. have undergone pretreatment, posttreatment and regular follow-up (at least annual) head and neck (H&N) magnetic resonance (MR) examinations.
6. have high-resolution H&N follow-up MR images, including axial T1- and T2-weighted imaging (T1-WI, T2-WI) and contrast-enhanced T1-weighted imaging (CE T1-WI).

5.1.2 Subject Exclusion Criteria

The subject must be excluded from participating in the trial if the subject:

1. had distant metastasis at the first diagnosis.
2. did not receive radical RT, or interruption time of RT > 7 days.
3. lacked T1-, T2-WI or CE T1-WI images or had pretreatment or follow-up MR images of insufficient quality to obtain measurements (i.e., motion artefacts).
4. had an unknown efficacy evaluation, partial remission (PR) or stable disease (SD) status beyond 6 months after RT ²⁵.
5. had irregular follow-up surveys.
6. had other malignancies or previous antitumor therapy.

5.1.3 Subject Withdrawal/Discontinuation Criteria

The subject will be withdrawn from participating in the trial if any of the following events occur during the trial:

1. The subject withdraws/discontinues from the regular follow-up survey (at least annually) during the trial.
2. The subject does not undergo H&N MR scanning at regular intervals (maximum interval between scans ≥ 24 months).

5.2 Basic Information Collection and Serial Number Generation

Once enrolled, the basic information, including demographics and baseline clinical characteristics of each subject, will be recorded by investigators as outlined in Table 1. A unique tracking number will be generated randomly from 0001 to 1 500 for each subject and be used to identify the subject for all procedures in the trial.

Table 1. Basic Information

Tracking No.	
Subject Name	
Enrollment Date	
Source of Center	
Gender	
Age	

Clinical Stage	
Treatment Regimen	
Date of Completion of Treatment	
Date of Complete Remission	
Dates of follow-up H&N MR surveys	
Recurrence or Non-recurrence	
Follow-up Duration (months)	

5.3 Imaging Data Collection and Process

5.3.1 Magnetic Resonance Imaging Collection and Anonymization

For each subject, initial tumor imaging by H&N MR scanning should have been performed within 1-2 weeks prior to treatment initiation. The process for imaging collection and transmission is manipulated in a uniform imaging protocol by radiologists and technicians in participating institutions. The whole series of MR scans should be exported as DICOM files and completely anonymized with unique tracking number before uploaded to the designated cloud platform in Hospital 1. Specially, the three sequences including axial high-resolution T1-WI, T2-WI, and CE T1-WI are requisite for the trial. The acquisition parameters of MRI scans derived from specific-vendor devices (1.0-, or 1.5-, or 3.0-T scanners) will be recorded as outlined in Table 2.

Table 2. Image Acquisition Parameters of MR Scans

Tracking No.	Scanner	Sequence	Magnetic field strength (T)	TR (ms)	TE (ms)	Pixel size (mm)	FOV (mm)	Slice thickness (mm)
Survey No.		T1-WI						
		T2-WI						
		CE T1-WI						

*Abbreviations: MR = magnetic resonance, T1-WI = T1-weighted imaging, T2-WI = T2-weighted imaging, CE T1-WI = contrast-enhanced T1-weighted imaging, TR = repetition time, TE = echo time, FOV = field of view.

5.3.2 Magnetic Resonance Imaging Quality Control

MR scans acquired at the sites will be downloaded and reviewed in Horos DICOM VIEWER (Version 3.3.6) or RadiAnt DICOM VIEWER (Version 5.0.1) by an independent radiologist experienced in H&N MR in the central laboratory to ensure high image quality for analysis. Cases that lack any requisite sequence (T1-WI, T2-WI, or CE T1-WI) will be excluded first. Images with insufficient clarity, low resolution, motion artefacts or other factors that might potentially affect imaging analysis will be further excluded.

5.3.3 Magnetic Resonance Imaging Labelling

All patients' follow-up H&N MR scans within the follow-up period are evaluated at the individual level by a rotating panel of four board-certified experts (two radiologists and two oncologists) from Hospital 1. The experts are free to indicate any suspected recurrence lesions on the MR images. Once a 'suspected' radiologically visible lesion was confirmed to be recurrence by clinical records, a labelling procedure is followed to determine the scan with the earliest visible recurrence (Table 3). Based on the typical and apparent recurrent lesions on the MR images at the diagnosis of recurrence, the MR images prior to that date are traced backtracks by the same group of experts to find the scan that showed the first signs of recurrence. Experts could consider all known clinical information to make a final decision.

Disagreements about the scans are arbitrated and the final decision is decided by a senior oncologist and a senior radiologist independently of the original four labellers, but with knowledge of their labels. The image labels are finalized only when the four experts from the same group reached a consensus.

Because other imaging modalities often used in clinical practice, such as nasopharyngeal endoscopy, H&N computed tomography, and whole-body positron emission tomography, were unavailable (as these are not routinely performed during surveillance), we describe the MR-derived early recurrence label as a silver standard. When the early recurrence scan identified by experts and the diagnostic recurrence scan overlapped, the label is defined as the scan with the earliest visible recurrence.

Table 3. Labeling Tool Prompts and Instructions

Patient ID:	Age:	Sex:	DOB:	Date of treatment completion:
Question				Possible answers (mark the best item with tick in questions with multiple-choices)
Has this patient achieved complete remission (CR) from (chemo)radiotherapy?				<u>Yes</u> : if there is no visible tumor in the nasopharynx (including the adjacent tissue and structure), bilateral retropharyngeal lymph node regions and bilateral cervical lymph node regions <u>Possibly*</u> : if there is suspected residual tumor in the nasopharynx (including the adjacent tissue and structure), bilateral retropharyngeal lymph node regions and bilateral cervical lymph node regions <u>No*</u> : if there is apparent residual tumor in the nasopharynx (including the adjacent tissue and structure), bilateral retropharyngeal lymph node regions and bilateral cervical lymph node regions
Does an event of recurrence occur in this patient within the follow-up period?				<u>Yes</u> : if there is any biopsy-, functional imaging-, or radiological evidence confirming recurrence in a local, regional or locoregional site <u>No*</u> : if there is no evidence of recurrence
Can you specify the site(s) of recurrence on the MR images?				<u>Local</u> : visible of recurrent tumor in the nasopharynx (including the adjacent tissue and structure) <u>Regional</u> : visible of recurrent tumor in the lateral or bilateral retropharyngeal lymph node region(s) <u>Locoregional</u> : visible recurrent tumor in the lateral or bilateral cervical lymph node region(s)
Please designate the date of the MR scan with the earliest visible recurrence:				
Confidence of designation of the scan with the earliest visible recurrence:				<u>5</u> : most certain about the condition <u>4</u> : <u>3</u> : <u>2</u> : <u>1</u> : least certain about the condition

*If these answers are selected, the remaining questions can be skipped.

5.4 Recurrence Detection by Model

The annotated images must be preprocessed and input into distinct prediction models to generate prediction values (varying from 0 to 1) for individuals following an established pipeline. The median value 0.5 is the threshold to classify positive labels or negative labels. Accordingly, individuals with prediction values higher than 0.5 will be defined as having ‘recurrence’, while those with values lower than 0.5 will be defined as having ‘nonrecurrence’. The prediction results will be completely anonymized with the corresponding tracking number and blinded to the subjects, treating physicians, and radiologists engaged in image labelling.

5.5 Manual Recurrence Detection

In a comparative study evaluating the performances of RAIN and humans in a prospective cohort, three radiologists from Hospital 1 will be recruited. Testers will be independent and will not be involved in labelling the scans. The testers will be informed that the scans to be read are of patients with treated NPC. The primary task will be to identify the first scan with recurrence. To simulate a realistic clinical workflow, scans will be presented chronologically to avoid revealing future scans ahead. When considering a scan (recurrence or nonrecurrence), testers will be allowed to refer to prior longitudinal MR scans, but not subsequent scan(s) (Fig. 2). Testers will be required to fill out the context questions listed in Table 3.

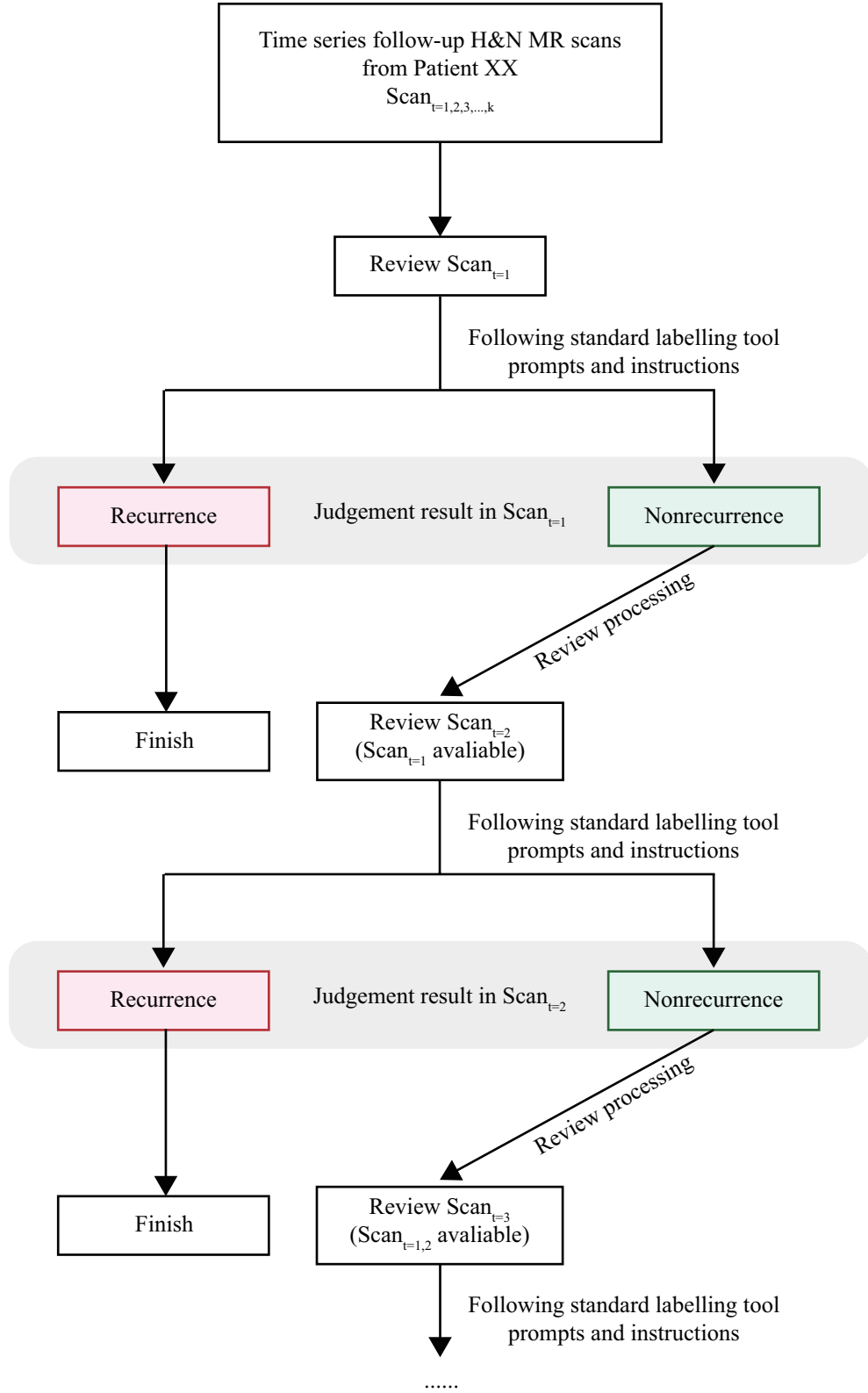


Figure 2. Scan Review Process

In prospective reader study, follow-up H&N MR scans from patients with treated NPC are presented to a rotating panel of readers at the individual level in chronological order. The readers independently review the scans. A follow-up H&N MR scan is labelled either positive for recurrence or negative for recurrence following standard labelling tool prompts and the instructions listed in Table 3.

Abbreviations: H&N = head and neck, MR = magnetic resonance, t = time.

5.6 Actual Endpoints

Treatment response will evaluate on each patient's H&N MR images within 6 months of the end of RT by four experts (two radiologists and two oncologists) specialized in NPC from Hospital 1. The maximum tumor diameter (MTD) of the primary tumor and metastatic lymph node(s) are separately measured on MR images. Tumor remission is evaluated based on the change in total MTD (estimated as the sum of the MTD of the primary tumor and metastatic lymph nodes) before and after treatment. Tumor remission is divided into 4 levels: complete remission (CR); partial remission (PR); stable disease (SD); and progressive disease (PD). Since almost all patients with NPC will achieve good disease control after radical (chemo)radiotherapy (87.0–97.0%)⁴⁻⁶, CR is defined as complete remission of the disease or the absence of any residual tumor, including both the primary tumor at the nasopharynx and the metastatic cervical lymph node(s), on MR images. PR is defined as the presence of residual disease in either the primary site and/or metastatic cervical lymph node(s). SD and PD are defined according to the Response Evaluation Criteria in Solid Tumors (RECIST, version 1.1)²⁵.

In cases of recurrence, for lesions that were accessible, the diagnosis is based on the positive histopathological results from biopsy (any component with cancerous cells or tissue originating from the primary NPC). For lesions that were not accessible, the diagnoses are made based on radiologic images. For patients with available functional 18F-FDG PET/CT imaging, the clinical diagnosis is made in consensus by two nuclear medicine physicians (each with 5 years of experience in PET/CT) using a GE Xeleris workstation. Image interpretation is based on visual evaluations. Any focus of FDG uptake greater than the surrounding background and not attributable to normal FDG biodistribution is assessed. The intensity of FDG uptake is graded using the five-point scale proposed by Ng et al^{26,27}. The probability of recurrent tumors is graded by using a five-point scoring system (0 = no lesion, 1 = definitely benign, 2 = probably benign, 3 = probably malignant, and 4 = definitely malignant). Grade 3 and grade 4 are both considered to be positive results. The sites of the positive results are recorded. In patients with a clinical diagnosis based on MR images, pre, posttreatment and “suspected recurrence” MR scans for each patient will be independently reviewed by a rotating panel of four board-certified experts (two radiologists and two oncologists) from Hospital 1.

Local recurrence is diagnosed when a focal area that showed the following was noted:

1. intermediate signal intensity on T1-WI;
2. enhancement on CE T1-WI;
3. high signal intensity on T2-WI images.

Regional recurrence is diagnosed when the following were observed:

1. an enlarged cervical lymph node (LN) with a minimal axial diameter (MID) ≥ 10 mm in the largest plane;
2. an enlarged lateral retropharyngeal LN with a MID ≥ 5 mm or any enlarged node in the median retropharyngeal group;
3. an LN of any size with central necrosis or a contrast-enhanced rim;
4. nodal grouping, the presence of ≥ 3 contiguous and confluent LNs, each with an MID 8–10 mm;
5. an LN of any size with extracapsular spread.

All enhanced lesions on MR images, irrespective of patient age, are considered tumor recurrence until proven otherwise. Fibrosis is diagnosed with confidence only if there is no contrast enhancement coupled with low signal intensity on T2-WI.

In cases of nonrecurrence within the follow-up duration, CR is defined as complete resolution of the

disease or absence of any residual tumor in both the primary and nodal sites on posttreatment MR scans, also including stable posttreatment changes seen on later imaging within a 5-year follow-up period.

The report will be documented as outlined in Table 4.

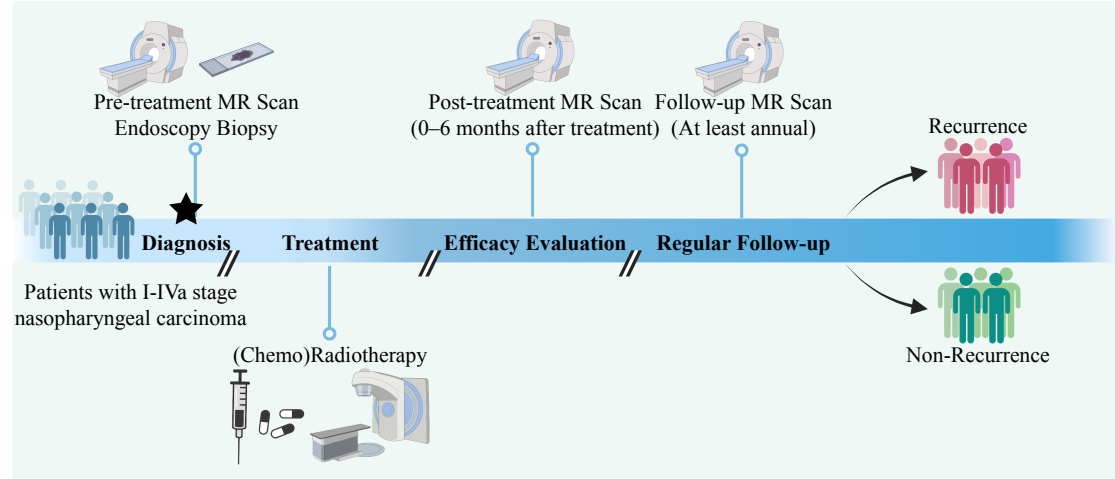
Table 4. Report of Survivor

Tracking No.	Survey Date	Recurrence or Non- recurrence	Recurrence Type (if recurrence)	r T Stage (if recurrence)	r N Stage (if recurrence)

6.0 TRIAL PROCEDURES

The trial procedures are summarized in the flow charts as described below. The asterisk under the time-course line indicates the possible timepoint a subject might be enrolled in the trial. The procedure will be performed accordingly.

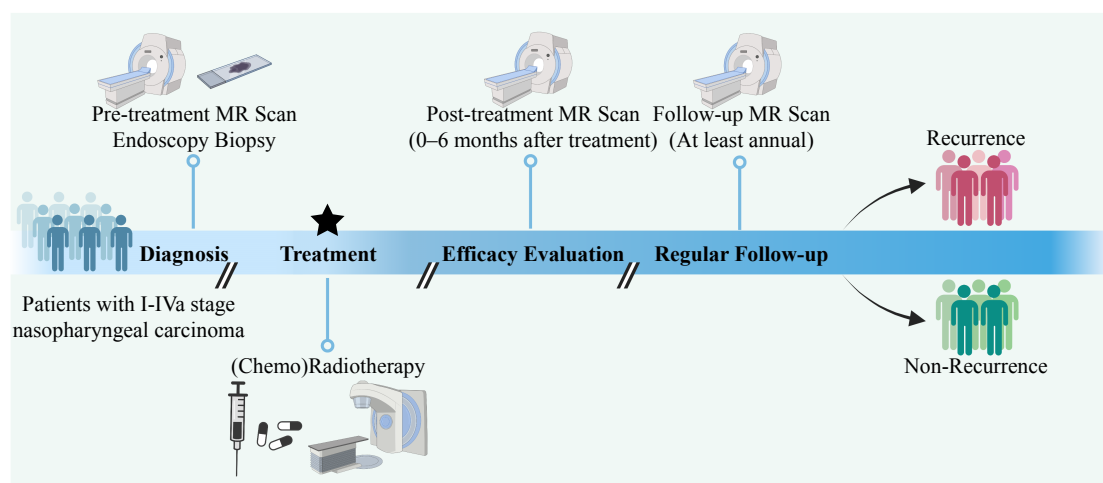
Flow chart 1. Subject is enrolled when he/she is newly diagnosed before treatment.



Trial Period	Diagnosis Phase	Treatment Phase	Efficacy Evaluation Phase	Regular Follow-up Phase
Administrative Procedures				
Informed Consent for Collection and Application of Clinical Sample and Medical Data	×			
Inclusion/Exclusion Criteria	×			
Subject Identification Tracking Number	×			
Demographics and Baseline Characteristics	×			
Image Data Collection and Analysis				
Pre-treatment MRI collection & labelling	×			
Post-treatment MRI collection & labelling			×	
Follow-up MRI collection & labelling				×
Trial Procedures				
RAIN Detection				×
Manual Reading	×	×	×	×

Follow-up				
Treatment Course Record		×		
Treatment Efficacy Record			×	
Actual Endpoints Record				×
Statistical Analysis				
Accuracy Analysis				×
Comparative Analysis				×

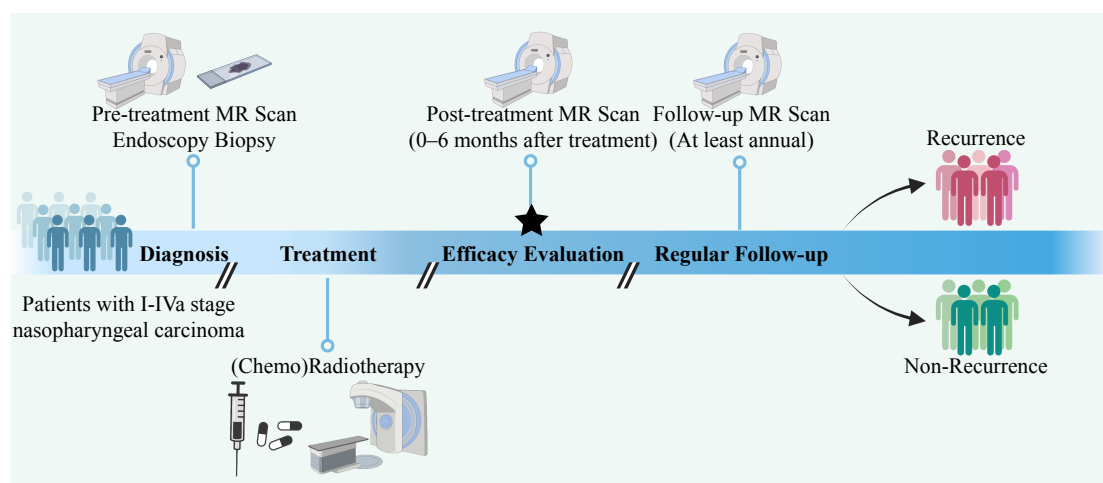
Flow chart 2. Subject is enrolled when he/she is undergoing (chemo)radiotherapy.



Trial Period	Diagnosis Phase	Treatment Phase	Efficacy Evaluation Phase	Regular Follow-up Phase
Administrative Procedures				
Informed Consent for Collection and Application of Clinical Sample and Medical Data	×			
Inclusion/Exclusion Criteria		×		
Subject Identification Tracking Number		×		
Demographics and Baseline Characteristics		×		
Image Data Collection and Analysis				
Pre-treatment MRI collection & labelling		×		
Post-treatment MRI collection & labelling			×	
Follow-up MRI collection & labelling				×
Trial Procedures				
RAIN Detection				×
Manual Reading	×	×	×	×
Follow-up				
Treatment Course Record		×		
Treatment Efficacy			×	

Record				
Actual Endpoints Record				×
Statistical Analysis				
Accuracy Analysis				×
Comparative Analysis				×

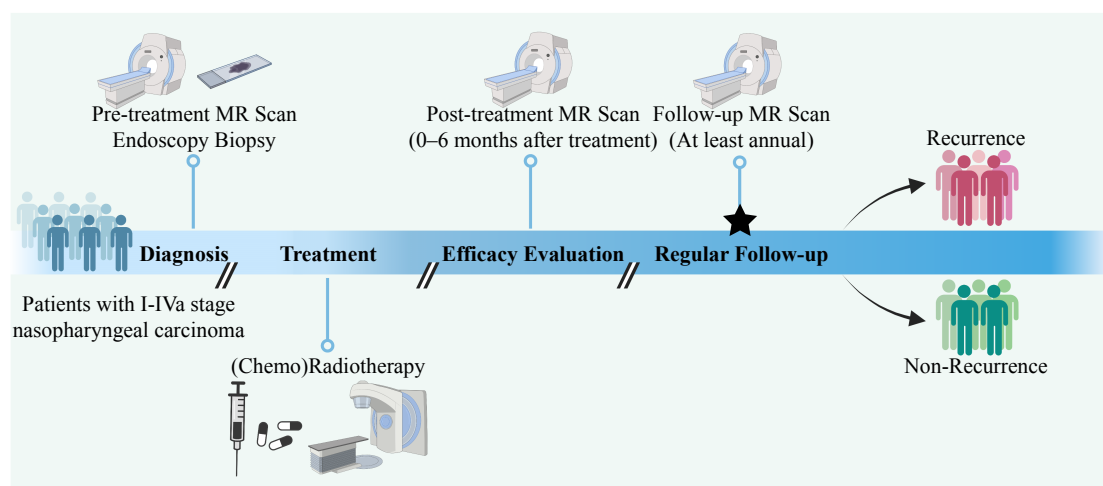
Flow chart 3. Subject is enrolled when he/she completed (chemo)radiotherapy.



Trial Period	Diagnosis Phase	Treatment Phase	Efficacy Evaluation Phase	Regular Follow-up Phase
Administrative Procedures				
Informed Consent for Collection and Application of Clinical Sample and Medical Data	×			
Inclusion/Exclusion Criteria			×	
Subject Identification Tracking Number			×	
Demographics and Baseline Characteristics			×	
Image Data Collection and Analysis				
Pre-treatment MRI collection & labelling			×	
Post-treatment MRI collection & labelling			×	
Follow-up MRI collection & labelling				×
Trial Procedures				
RAIN Detection				×
Manual Reading	×	×	×	×
Follow-up				
Treatment Course Record			×	
Treatment Efficacy			×	

Record				
Actual Endpoints Record				×
Statistical Analysis				
Accuracy Analysis				×
Comparative Analysis				×

Flow chart 4. Subject is enrolled when he/she is undergoing regular follow-up.



Trial Period	Diagnosis Phase	Treatment Phase	Efficacy Evaluation Phase	Regular Follow-up Phase
Administrative Procedures				
Informed Consent for Collection and Application of Clinical Sample and Medical Data	×			
Inclusion/Exclusion Criteria				×
Subject Identification Tracking Number				×
Demographics and Baseline Characteristics				×
Image Data Collection and Analysis				
Pre-treatment MRI collection & labelling				×
Post-treatment MRI collection & labelling				×
Follow-up MRI collection & labelling				×
Trial Procedures				
RAIN Detection				×
Manual Reading	×	×	×	×
Follow-up				
Treatment Course Record				×
Treatment Efficacy				×

Record				
Actual Endpoints Record				×
Statistical Analysis				
Accuracy Analysis				×
Comparative Analysis				×

7.0 STATISTICAL ANALYSIS

This section outlines the statistical strategy and procedures for the trial. If, after the trial has begun, changes are made to the primary and/or key secondary hypotheses, or the statistical methods related to those hypotheses, then the protocol will be amended accordingly.

7.1 Statistical Analysis Plan Summary

Study Design Overview	A multicenter, observational diagnostic study for the development and retrospective and prospective validation of a deep learning (DL)-based recurrence detection model (Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma, RAIN) for diagnosing the earliest radiologically visible recurrence on follow-up MR images of patients with treated nonmetastatic NPC.
Detection Assignment	Approximately 1 500 patients will be retrospectively enrolled from two centers into a development and internal and external validation cohorts. Approximately 245 patients will be prospectively enrolled in Hospital 1 as a prospective validation cohort. All follow-up MR images from patients with CR of NPC in the validation cohorts will be evaluated as showing 'recurrence' or 'nonrecurrence' by the RAIN detector (See Section 5.4 for details). Follow-up MR images from patients with CR of NPC in the prospective validation cohort will be evaluated as showing 'recurrence' or 'nonrecurrence' by three readers independently.
Analysis Populations	Accuracy: All validation cohorts. Comparative: All validation cohorts.
Primary Endpoint	AUC
Statistical Methods for Key Accuracy Analyses	The binary prediction results generated by RAIN will be compared to the ground truth labelled by four experts, generating a prediction matrix. The primary hypothesis for AUC (H1) will be evaluated by calculating the area under the receiver operating characteristics (ROC) curve which plots the proportion of true-positive cases (sensitivity) against the proportion of false-positive cases (1-specificity) based on various predictive probability thresholds. The 95% confidence intervals (95% CI) will be generated through bootstrapping with 1 000 sampling trials.
Statistical Methods for Key Comparative	The AUCs and 95% CIs of the RAIN detector and

Analyses	three readers will be calculated similarly based on the ground truth. The secondary hypothesis (H2) will be evaluated through pairwise comparisons of AUCs between the RAIN detector and three readers. Delong's test or alternatively, Student's t test, will be performed to obtain a p value indicating whether RAIN has a significantly different AUC from the three readers.
Sample Size and Power	Due to the retrospective nature of our study in developing the DL-based model, a total of 1 500 subjects were eligible for development and retrospective validation for the model. The planned sample size in the prospective validation cohort is approximately 245 subjects. This study has a diagnostic test design. The primary objective is the area under the curve (AUC). The single ROC sample size estimation module of Power Analysis and Sample Size (PASS) software was adopted. The relevant parameters were set as follows: alpha (α) = 0.05 (two-sided), power ($1-\beta$) = 85%, positive cases presumably accounting for 25% of the population, null hypothesis AUC = 0.72, expected AUC = 0.85. A total of 49 positive cases and 147 negative cases should be included in the population. Considering that a loss to follow-up rate of 20%, 61 positive cases and 184 negative cases should be included in the population.

7.2 Hypotheses/Estimation

The objectives and hypotheses of the study are stated in Section 3.0.

7.3 Analysis Population

Eligible subjects in all validation cohorts will serve as the population for both the primary accuracy analysis and secondary comparative and accuracy analyses. All enrolled subjects who undergo prospective evaluation by the RAIN detector with a final actual endpoint will be included in this population.

7.4 Statistical Methods for Study Endpoints

This section describes the statistical methods that address the primary and secondary objectives.

7.4.1 Primary Accuracy Endpoints

7.4.1.1 Area under curve (AUC)

AUC is defined as the probability that a randomly chosen positive example is ranked higher than a randomly chosen negative example. A higher AUC indicates a better classification performance of a definite predictor.

A prediction matrix will be generated by comparing the binary prediction results of RAIN with the ground truth labelled by four experts, thus indicating the distribution profile of true-positive, true-

negative, false-positive and false-negative cases. The AUC will be evaluated by calculating the area under the receiver operating characteristic (ROC) curve, which plots the proportion of true-positive cases (sensitivity) against the proportion of false-positive cases (1-specificity) based on various predictive probability thresholds. The 95% confidence intervals (95% CI) of the AUCs will be generated through bootstrapping with 1 000 sampling trials.

7.4.2 Secondary Accuracy Endpoints

7.4.2.1 Sensitivity

Sensitivity is defined as the proportion of the predicted positive cases among the total actual positive cases, also known as the true-positive rate.

In this study, the sensitivities of RAIN and the readers will be evaluated by calculating the proportion of subjects with ‘predicted recurrence’ among the total subjects with ‘actual recurrence’. The 95% CIs of sensitivity will be generated through bootstrapping with 1 000 sampling trials.

7.4.2.2 Specificity

Specificity is defined as the proportion of the predicted negative cases among the total actual negative cases, also known as the true-negative rate.

In this study, the specificities of RAIN and the readers will be evaluated by calculating the proportion of subjects with ‘predicted nonrecurrence’ among the total subjects with ‘actual nonrecurrence’. The 95% CI of specificity will be generated through bootstrapping with 1 000 sampling trials.

7.4.2.3 Positive Prediction Value

Positive prediction value (PPV) is defined as the proportion of the actual positive cases among the total predicted positive cases.

In this study, the PPVs of RAIN and the readers will be evaluated by calculating the proportion of subjects with ‘actual recurrence’ among the total subjects with ‘predicted recurrence’. The 95% CIs of the PPV will be generated through bootstrapping with 1 000 sampling trials.

7.4.2.4 Negative Prediction Value

Negative prediction value (NPV) is defined as the proportion of the actual negative cases among the total predicted negative cases.

In this study, the NPVs of RAIN and the reader will be evaluated by calculating the proportion of subjects with ‘actual nonrecurrence’ among the total subjects with ‘predicted nonrecurrence’. The 95% CI of the NPV will be generated through bootstrapping with 1 000 sampling trials.

7.4.3 Secondary Comparative Endpoints

7.4.3.1 Delong’s test

Delong’s test is a nonparametric test for comparing the AUCs of two or more correlated ROC curves.

The AUC of RAIN will be compared to the AUCs of the other three readers by Delong’s test. A two-sided p value of less than 0.05 was considered to indicate a significant difference.

7.4.3.2 Student’s t test

Student’s t test is an inferential statistic used to determine if there is a significant difference between the means of two datasets that follow normal distributions with homogeneous variance.

The mean AUC of RAIN will be compared to the mean AUCs of the other three readers by Student’s t test. A two-sided p value of less than 0.05 was considered to indicate a significant difference.

7.4.3.3 McNemar test

McNemar test is a test on a 2×2 classification table to test the difference between paired proportions.

The sensitivity, specificity, and accuracy of the RAIN model and human readers will be compared by using the McNemar test. A two-sided p value of less than 0.05 was considered to indicate a significant

difference.

7.5 Statistical Methods for Baseline Characteristics and Demographics

The comparability of each relevant baseline characteristic between the two groups with distinct endpoints (recurrence *vs* nonrecurrence) will be assessed by the use of tables and/or graphs. The number and percentage of subjects in each subgroup will be displayed. Demographic variables (such as age and sex) and baseline characteristics (such as clinical stage) will be summarized either by descriptive statistics or categorical tables. Differences in nonnormally distributed variables between groups were examined with the Mann-Whitney U test, while χ^2 test or Fisher's exact test will be used for categorical variables.

7.6 Sample Size and Power Calculations

Due to the retrospective nature of our study in developing the DL-based model, a total of 1 500 subjects are eligible for development and retrospective validation for the model. The planned sample size in the prospective validation cohort is approximately 245 subjects. This study has a diagnostic test design. The primary objective is the area under the curve (AUC). The single ROC sample size estimation module of Power Analysis and Sample Size (PASS) software was adopted. The relevant parameters were set as follows: alpha (α) = 0.05 (two-sided), power ($1-\beta$) = 85%, positive cases presumably accounting for 25% of the population, null hypothesis AUC = 0.72, and expected AUC = 0.85. A total of 49 positive cases and 147 negative cases should be included in the population. Considering that a loss to follow-up rate of 20%, 61 positive cases and 184 negative cases should be included in the population.

8.0 CONFIDENTIALITY AND DATA SHARING PLAN

8.1 Confidentiality of Data

Data generated in the trials will be considered confidential by the investigators, except to the extent that they are included in a publication. During the trial, each subject will be identified by a unique tracking number.

8.2 Data Sharing Plan

The subject data (deidentified participant information and original MR images) and the full study protocol will be made available to the scientific community immediately upon publication, with as few restrictions as possible. All requests should be submitted to the investigators for consideration. A data use agreement will be required before the release of subject data and approval from the institutional review board, as appropriate. The source code will be openly available on the website until publication.

9.0 SPONSORS AND COLLABORATORS

9.1 Sponsors and Collaborators

The trial is sponsored by Hospital 1. The Study Protocol is completed and reviewed by all the participating institutions.

9.2 Responsible Party/Investigator

Principal Investigators are responsible for the study protocol. The central contact information is as following:

XXX (Blinded Name)

Telephone: XXX (Blinded Number)

Email: (Blinded address)

9.3 Role of Funding

The study protocol, design and data requirements were approved and sponsored by XXX.

The funders have no role in the study design, data collection, data analysis, data interpretation, writing of the report, or the decision to submit the report for publication.

9.4 Ethics Committee

This study was approved by the independent institutional review board of Hospital 1 (Blinded Approved ID) and performed according to the Helsinki declaration.

10.0 REFERENCES

1. Shapiro, C. L. Cancer Survivorship. *N. Engl. J. Med.* **379**, 2438–2450 (2018).
2. Mahvi, D. A., Liu, R., Grinstaff, M. W., Colson, Y. L. & Raut, C. P. Local Cancer Recurrence: The Realities, Challenges, and Opportunities for New Therapies. *CA. Cancer J. Clin.* **68**, 488–505 (2018).
3. Siegel, R. L., Miller, K. D., Fuchs, H. E. & Jemal, A. Cancer Statistics, 2021. *CA. Cancer J. Clin.* **71**, 7–33 (2021).
4. Lee, A. W. M. *et al.* Evolution of treatment for nasopharyngeal cancer – Success and setback in the intensity-modulated radiotherapy era. *Radiother. Oncol.* **110**, 377–384 (2014).
5. Zhang, M.-X. *et al.* Intensity-modulated radiotherapy prolongs the survival of patients with nasopharyngeal carcinoma compared with conventional two-dimensional radiotherapy: A 10-year experience with a large cohort and long follow-up. *Eur. J. Cancer* **51**, 2587–2595 (2015).
6. Mao, Y.-P. *et al.* Prognostic factors and failure patterns in non-metastatic nasopharyngeal carcinoma after intensity-modulated radiotherapy. *Chin. J. Cancer* **35**, 103 (2016).
7. Chang, J. T. *et al.* Locally recurrent nasopharyngeal carcinoma. *Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol.* **54**, 135–142 (2000).
8. Yeung, D. C. M., Yeung, Z., Wong, E. W. Y., Vlantis, A. C. & Chan, J. Y. K. Neck lymph node status on survival of regionally recurrent or persistent nasopharyngeal carcinoma. *Sci Rep* **10**, 5622 (2020).
9. Hu, J. *et al.* Clinical outcomes of carbon-ion radiotherapy for patients with locoregionally recurrent nasopharyngeal carcinoma. *Cancer* **126**, 5173–5183 (2020).
10. Pfister, D. G. *et al.* Head and Neck Cancers, Version 2.2020, NCCN Clinical Practice Guidelines in Oncology. *J. Natl. Compr. Cancer Netw. JNCCN* **18**, 873–898 (2020).
11. Tang, L.-L. *et al.* The Chinese Society of Clinical Oncology (CSCO) clinical guidelines for the diagnosis and treatment of nasopharyngeal carcinoma. *Cancer Commun. Lond. Engl.* **41**, 1195–1227

- (2021).
12. Meng, K., Tey, J., Ho, F. C. H., Asim, H. & Cheo, T. Utility of magnetic resonance imaging in determining treatment response and local recurrence in nasopharyngeal carcinoma treated curatively. *BMC Cancer* **20**, 193 (2020).
 13. Fitzgerald, R. Error in radiology. *Clin. Radiol.* **56**, 938–946 (2001).
 14. Yala, A. *et al.* Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat. Med.* **28**, 136–143 (2022).
 15. Zhang, B. *et al.* Radiomics Features of Multiparametric MRI as Novel Prognostic Factors in Advanced Nasopharyngeal Carcinoma. *Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res.* **23**, 4259–4269 (2017).
 16. Zhang, L.-L. *et al.* Pretreatment MRI radiomics analysis allows for reliable prediction of local recurrence in non-metastatic T4 nasopharyngeal carcinoma. *EBioMedicine* **42**, 270–280 (2019).
 17. Dong, D. *et al.* Development and validation of a novel MR imaging predictor of response to induction chemotherapy in locoregionally advanced nasopharyngeal cancer: a randomized controlled trial substudy (NCT01245959). *BMC Med.* **17**, 190 (2019).
 18. Qiang, M. *et al.* A Prognostic Predictive System Based on Deep Learning for Locoregionally Advanced Nasopharyngeal Carcinoma. *J. Natl. Cancer Inst.* **113**, 606–615 (2021).
 19. Gulshan, V. *et al.* Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA* **316**, 2402–2410 (2016).
 20. Bi, W. L. *et al.* Artificial intelligence in cancer imaging: Clinical challenges and applications. *CA. Cancer J. Clin.* **69**, 127–157 (2019).
 21. Esteva, A. *et al.* A guide to deep learning in healthcare. *Nat. Med.* **25**, 24–29 (2019).

22. Yu, K.-H., Beam, A. L. & Kohane, I. S. Artificial intelligence in healthcare. *Nat. Biomed. Eng.* **2**, 719–731 (2018).
23. Castellino, R. A. Computer aided detection (CAD): an overview. *Cancer Imaging Off. Publ. Int. Cancer Imaging Soc.* **5**, 17–19 (2005).
24. Amin, M. B. *et al.* The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more ‘personalized’ approach to cancer staging. *CA. Cancer J. Clin.* **67**, 93–99 (2017).
25. Therasse, P. *et al.* New Guidelines to Evaluate the Response to Treatment in Solid Tumors. *JNCI J. Natl. Cancer Inst.* **92**, 205–216 (2000).
26. Ng, S.-H. *et al.* Nodal metastases of nasopharyngeal carcinoma: patterns of disease on MRI and FDG PET. *Eur. J. Nucl. Med. Mol. Imaging* **31**, 1073–1080 (2004).
27. Ng, S.-H. *et al.* Comprehensive imaging of residual/ recurrent nasopharyngeal carcinoma using whole-body MRI at 3 T compared with FDG-PET-CT. *Eur. Radiol.* **20**, 2229–2240 (2010).