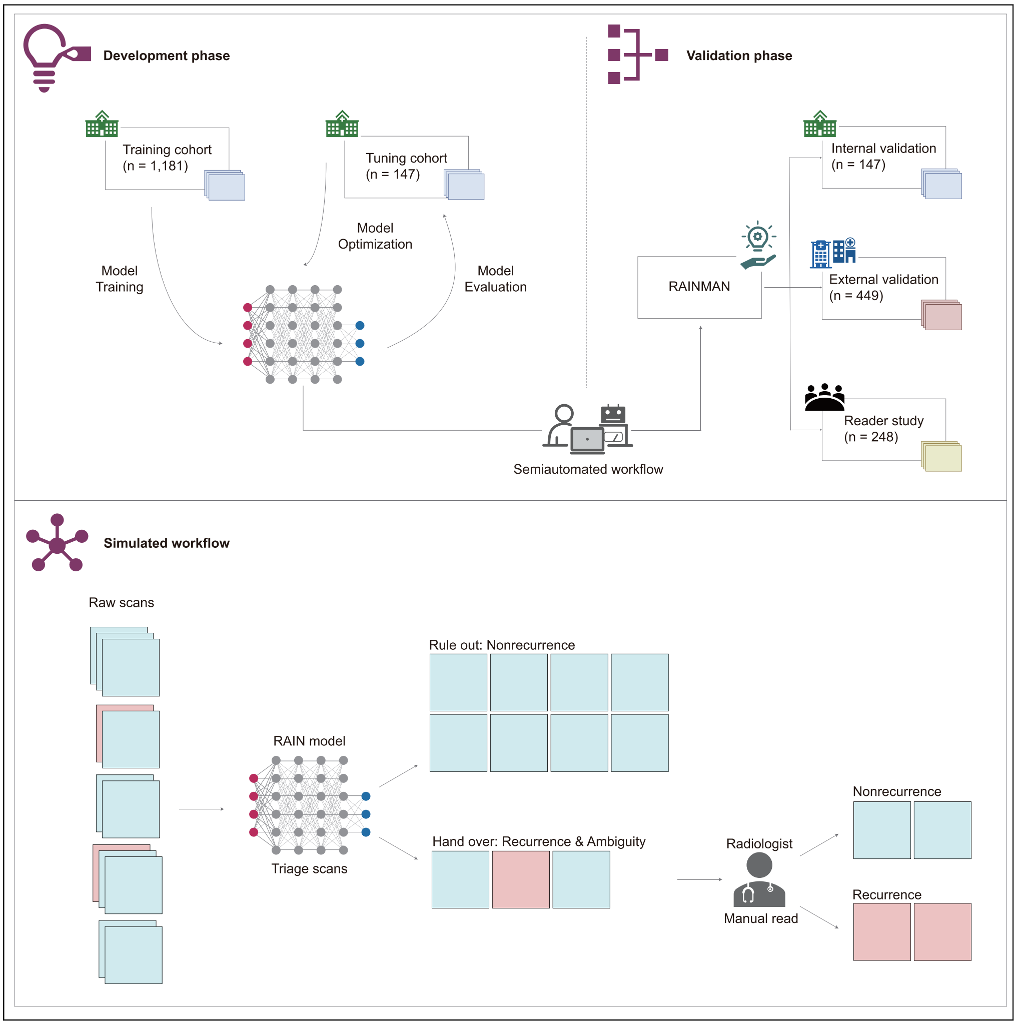
iScience

CellPress
OPEN ACCESS

Article

# A deep learning-based semiautomated workflow for triaging follow-up MR scans in treated nasopharyngeal carcinoma



Ying-Ying Huang, Yi-Shu Deng, Yang Liu, ..., Liang-Ru Ke, Xing Lv, Chao-Feng Li

guoxiang@sysucc.org.cn (X.G.)
kelr@sysucc.org.cn (L.-R.K.)
lvxing@sysucc.org.cn (X.L.)
lichaofeng@sysucc.org.cn (C.-F.L.)

## Highlights

A semiautomated workflow was constructed for monitoring recurrent NPC on MR scans

The workflow reduces radiologist workload while retaining undamaged performance

The workflow provided an approach utilizing virtues of machine and human

Huang et al., iScience *26*, 108347
December 15, 2023 © 2023
https://doi.org/10.1016/j.isci.2023.108347

iScience

CellPress
OPEN ACCESS

Article

# A deep learning-based semiautomated workflow for triaging follow-up MR scans in treated nasopharyngeal carcinoma

Ying-Ying Huang,[1,2,10] Yi-Shu Deng,[1,3,4,10] Yang Liu,[1,5,10] Meng-Yun Qiang,[6,10] Wen-Ze Qiu,[7] Wei-Xiong Xia,[1,8] Bing-Zhong Jing,[1,3] Chen-Yang Feng,[1,3] Hao-Hua Chen,[1,3] Xun Cao,[1,9] Jia-Yu Zhou,[1,8] Hao-Yang Huang,[1,8] Ze-Jiang Zhan,[1,8] Ying Deng,[1,8] Lin-Quan Tang,[1,8] Hai-Qiang Mai,[1,8] Ying Sun,[1,5] Chuan-Miao Xie,[1,2] Xiang Guo,[1,8,*] Liang-Ru Ke,[1,2,*] Xing Lv,[1,8,*] and Chao-Feng Li[1,3,11,*]

## SUMMARY

**It is imperative to optimally utilize virtues and obviate defects of fully automated analysis and expert knowledge in new paradigms of healthcare. We present a deep learning-based semiautomated workflow (RAINMAN) with 12,809 follow-up scans among 2,172 patients with treated nasopharyngeal carcinoma from three centers (ChiCTR.org.cn, Chi-CTR2200056595). A boost of diagnostic performance and reduced workload was observed in RAINMAN compared with the original manual interpretations (internal vs. external: sensitivity, 2.5% [p = 0.500] vs. 3.2% [p = 0.031]; specificity, 2.9% [p < 0.001] vs. 0.3% [p = 0.302]; workload reduction, 79.3% vs. 76.2%). The workflow also yielded a triaging performance of 83.6%, with increases of 1.5% in sensitivity (p = 1.000) and 0.6%–1.3% (all p < 0.05) in specificity compared to three radiologists in the reader study. The semiautomated workflow shows its unique superiority in reducing radiologist's workload by eliminating negative scans while retaining the diagnostic performance of radiologists.**

## INTRODUCTION

Long-term surveillance in disease control is pivotal in managing cancer patients.[1,2] Nasopharyngeal carcinoma (NPC), a cancer that prevails across East and Southeast Asia,[3] has a 5-year overall survival (OS) of 87.4%, and a locoregional recurrence rate of 7.4%–15.0%.[4–6] The annual proceeding of head and neck (H&N) magnetic resonance (MR) imaging has been a dominant follow-up scheme in monitoring locoregional disease control in NPC.[7,8] However, the interpretation of overwhelming imaging data involves the consumption of time and labor by relatively scarce radiologists.[9] Moreover, owing to histological and structural changes induced by radiation and disturbance of personal subjective judgment, even experienced radiologists may inevitably neglect some occult and atypical recurrence or waver for benign changes, resulting in mis- and missed diagnoses.[10,11] False-positive scans can bring about superfluous procedures, placing both psychological and economic burdens on well-healed patients.[10,11] Although double reading has slightly improved radiologist performance, it undeniably aggravates inefficiency and potentially increases false-positive results.[12] Therefore, a triaging workflow to eliminate the true negatives for which radiologist assessment is unnecessary and to distinguish the scans showing high likelihood of recurrence is needed.

An accumulation of advancements in deep learning (DL) engineered for computer-aided detection (CAD) in medical imaging have repeatedly been considered to save time and even improve radiologists' performance in the assessment of cancer screening.[13–20] Among them, especially many studies have introduced successful CAD tool for identifying breast cancer. They have reported workload reduction of 19.3%–72.5% at CAD tool on digital breast examinations.[16,17] However, a major drawback of some of these studies is that the radiologist still needs to evaluate all examinations, and thus the workload is not alleviated. Recently, inspiring work has evaluated

[1] State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, Guangzhou 510060, China
[2] Department of Radiology, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[3] Department of Information, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[4] School of Electronics and Information Technology, Sun Yat-sen University, Guangzhou 510006, China
[5] Department of Radiation Oncology, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[6] Department of Radiation Oncology, Cancer Hospital of The University of Chinese Academy of Sciences, Hangzhou 310005, China
[7] Department of Radiation Oncology, The Affiliated Cancer Hospital of Guangzhou Medical University, Guangzhou 510095, China
[8] Department of Nasopharyngeal Carcinoma, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[9] Department of Critical Care Medicine, Sun Yat-sen University Cancer Center, Guangzhou 510060, China
[10] These authors contributed equally
[11] Lead contact
*Correspondence: guoxiang@sysucc.org.cn (X.G.), kelr@sysucc.org.cn (L.-R.K.), lvxing@sysucc.org.cn (X.L.), lichaofeng@sysucc.org.cn (C.-F.L.)
https://doi.org/10.1016/j.isci.2023.108347

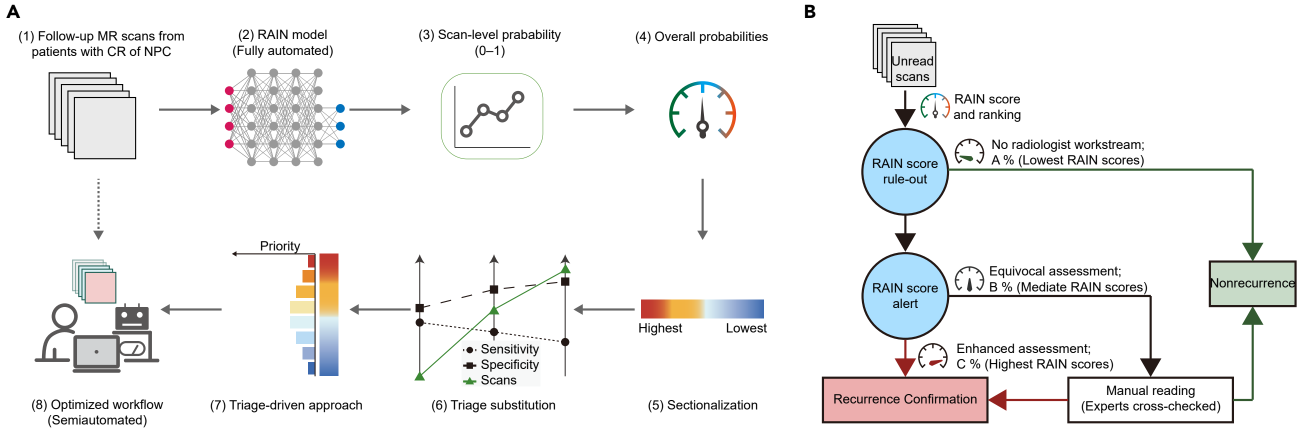

A
(1) Follow-up MR scans from patients with CR of NPC
(2) RAIN model (Fully automated)
(3) Scan-level probability (0–1)
(4) Overall probabilities
(5) Sectionalization
Highest
Lowest
(6) Triage substitution
Sensitivity
Specificity
Scans
(7) Triage-driven approach
Priority
(8) Optimized workflow (Semiautomated)
B
Unread scans
RAIN score and ranking
RAIN score rule-out
No radiologist workstream; A % (Lowest RAIN scores)
RAIN score alert
Equivocal assessment; B % (Mediate RAIN scores)
Enhanced assessment; C % (Highest RAIN scores)
Nonrecurrence
Manual reading (Experts cross-checked)
Recurrence Confirmation

**Figure 1. Study overview**

(A) Study flow chart.

(B) Proposed workflow for preselecting follow-up H&N MR scans among patients with CR of NPC. Abbreviations: H&N, head and neck; MR, magnetic resonance; CR, complete remission; NPC, nasopharyngeal carcinoma; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma.

value of DL in triaging medical images, whereby "high suspicious negative" examinations are remitted and the remaining examinations are further reviewed manually.[18–20] Although sizable reductions in the radiologist workload would be achieved, damaged sensitivity might come with it. A fully automated model surely frees up all radiologists, but computer-human collaboration that well balances the advantages of fully automated analysis and manual reading would be more feasible and promising. The strength of DL is the potential capacity of an assisted DL-based CAD tool to conserve resources and reduce oversights in the surveillance process among cancer survivors with treated NPC. Although there have been several studies assessing the stand-alone performance of CAD tool in detecting primary NPC tumor,[21–24] however, to date, there are no studies investigating DL-powered triaging approach in optimizing surveillance in survivors with cured NPC.

Here, we present a semiautomated triaging workflow (RAIN + MANual reading, RAINMAN) in a simulation analysis from a multicenter study (Figure 1). The RAINMAN workflow triages massive follow-up scans into three tiers as the detector first dismisses the majority of normal scans (no radiologist workstream) and singles out the high-confidence positives for enhanced assessment (enhanced assessment workstream), leaving the equivocal remainders for manual read. We envision that this computer-human collaboration will simplify surveillance among cancer patients.

## RESULTS

### Patient and scan characteristics

A total of 2,172 patients (mean age, 45.2 ± 10.3 years [standard deviation, SD]; 1,592 men) were enrolled in this study between September 10, 2007, and May 20, 2021 (Figure 2). Among them, 1,050 patients had confirmed recurrences. Clinical characteristics are listed in Table 1.

A total of 12,809 follow-up H&N MR scans were collected. Routine follow-up comprised annual MR inspection (mean interval, 8.0 ± 4.2 months [SD]) for monitoring locoregional disease reactivation (Table S1). A total of 8.2% (1,050 of 12,809) of MR scans were confirmed to have recurrence (Table S2).

### Retrospective study results

The RAIN model had an area under the receiver operating characteristic curve (ROC-AUC) of 0.982 (95% confidential interval [CI] 0.974–0.990) on tuning cohort (Figure S1). In the validation cohorts, the RAIN obtained ROC-AUCs of 0.990 (95% CI 0.984–0.995) internally and 0.958 (95% CI 0.950–0.966) externally (Figure 3). Further analysis for the performance of the RAIN model in binary classification task is available in Table S3.

We determined a rule-out threshold that maintains undamaged sensitivity with the 80% lowest RAIN scores and an alert threshold that maintains undamaged specificity with the 3% highest RAIN scores on tuning cohort (Figure 4). Alternative choices of triage threshold are available in Table S4.

When 100% of the scans were manually read, the original radiologists (according to documented radiologic reports) obtained an ROC-AUC of 0.927 (95% CI 0.891–0.962), with a sensitivity of 88.6% (95% CI 81.4–95.8) and a specificity of 96.7% (95% CI 95.4–98.0) within internal validation cohort (Figure 3 and Table 2). In the internal validation cohort, 88.3% of unequivocal negatives were substituted by RAIN (no radiologist workstream), resulting in a workload reduction of 79.3% and manual reading of the rest (20.7% equivocal scans) (Figure 3 and Table 2). In the enhanced assessment workstream, there were 34.2% scan-detected recurrences preselected by the RAIN detector. This triage-driven, semiautomated analysis would thus result both in a performance boost (ROC-AUC 0.954 [95% CI 0.922–0.985]; sensitivity 91.1%

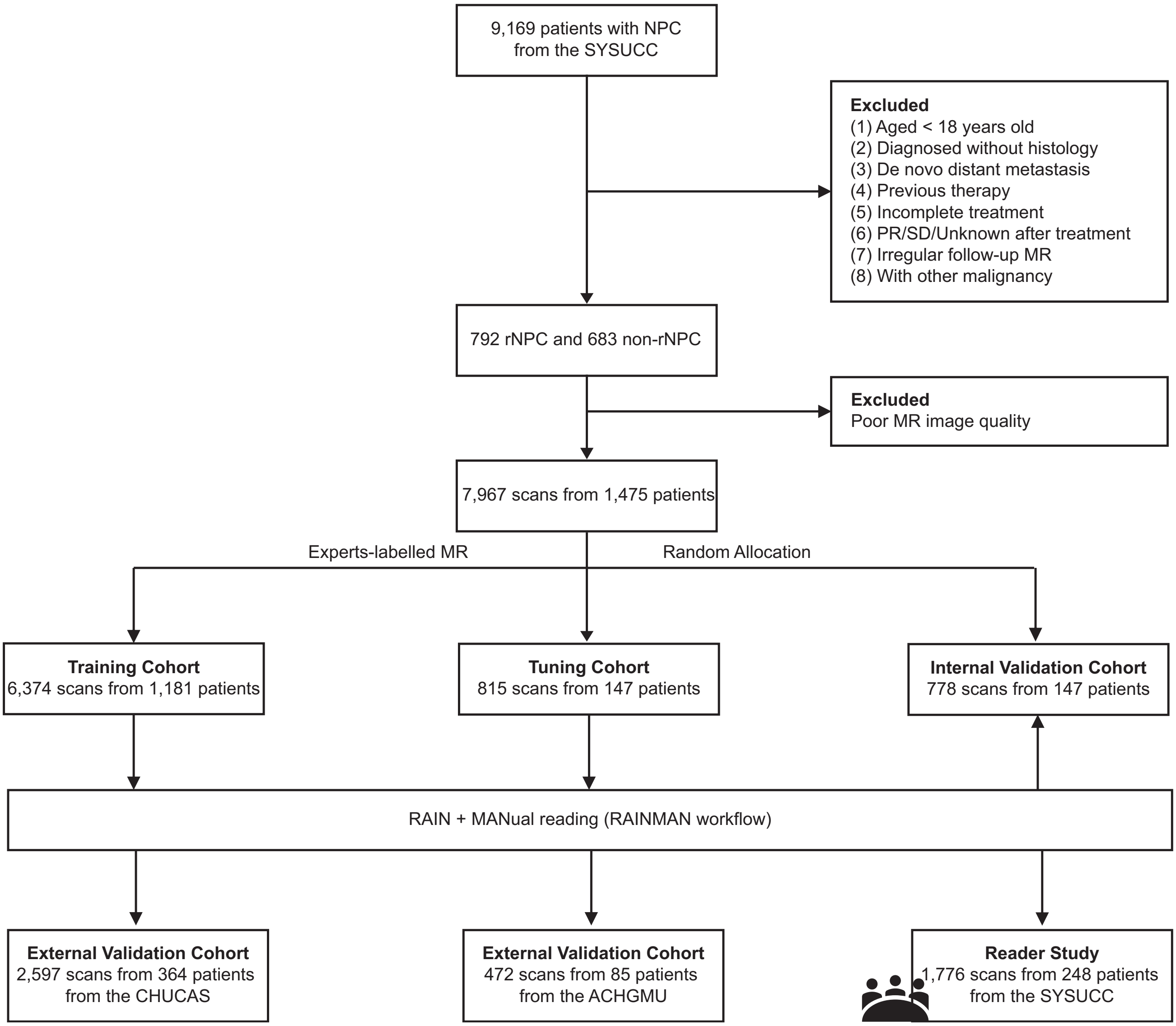

9,169 patients with NPC
from the SYSUCC
Excluded
(1) Aged < 18 years old
(2) Diagnosed without histology
(3) De novo distant metastasis
(4) Previous therapy
(5) Incomplete treatment
(6) PR/SD/Unknown after treatment
(7) Irregular follow-up MR
(8) With other malignancy
792 rNPC and 683 non-rNPC
Excluded
Poor MR image quality
7,967 scans from 1,475 patients
Experts-labelled MR
Random Allocation
Training Cohort
6,374 scans from 1,181 patients
Tuning Cohort
815 scans from 147 patients
Internal Validation Cohort
778 scans from 147 patients
RAIN + MANual reading (RAINMAN workflow)
External Validation Cohort
2,597 scans from 364 patients
from the CHUCAS
External Validation Cohort
472 scans from 85 patients
from the ACHGMU
Reader Study
1,776 scans from 248 patients
from the SYSUCC

**Figure 2. Study population flow chart**

Abbreviations: NPC, nasopharyngeal carcinoma; PR, partial remission; SD, stable disease; rNPC, recurrent nasopharyngeal carcinoma; non-rNPC, non-recurrent nasopharyngeal carcinoma; MR, magnetic resonance; SYSUCC, Sun Yat-sen University Cancer Center; CHUCAS, Cancer Hospital of The University of Chinese Academy of Sciences; ACHGMU, The Affiliated Cancer Hospital of Guangzhou Medical University.

[95% CI 84.7–97.5; p = 0.500]; specificity 99.6% [95% CI 99.1–100.0; p < 0.001]) and 79.3% labor saving by helping radiologists to concentrate on equivocal scans while leaving unequivocal negatives for fully automated analysis (Figure 3A and Table 2).

In the external validation cohort, the RAIN detector preselected 29.8% positives and eliminated 81.3% negatives. Thus, the simulated application of the RAINMAN workflow resulted in a workload reduction of 76.2%, a promotion in ROC-AUC of 0.935 (95% CI 0.912–0.957), and a sensitivity of 88.5% (95% CI 83.9–93.1) for recurrence detection when compared to radiologists (original radiologists: ROC-AUC: 0.917 [95% CI 0.892–0.942], sensitivity: 85.3% [95% CI 80.3–90.4]). The semiautomated triaging strategy also resulted in an inferior specificity of 98.4% (95% CI 98.0–98.9) to that of radiologists (98.1% [95% CI 97.6–98.6]) (Figure 3B and Table 2).

## Reader study results

The generalizability of RAIN was strengthened in a real-world screening cohort. The RAIN yielded an ROC-AUC of 0.969 (95% CI 0.960–0.979), with sensitivity of 91.0% (95% CI 84.0–8.1) and specificity of 91.7% (95% CI 90.4–93.0) in identifying scans with recurrence (Figure 5A and Table 2). Further analysis for the performance of the RAIN model in binary classification task is available in Table S3.

Table 1. Patient characteristics

| | Training (n = 1,181) | | Tuning (n = 147) | | Internal validation (n = 147) | | External validation (n = 449) | | Reader study (n = 248) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | rNPC (n = 634) | non-rNPC (n = 547) | rNPC (n = 79) | non-rNPC (n = 68) | rNPC (n = 79) | non-rNPC (n = 68) | rNPC (n = 191) | non-rNPC (n = 258) | rNPC (n = 67) | non-rNPC (n = 181) |
| Sex | | | | | | | | | | |
| Male | 481 (75.9) | 394 (72.0) | 57 (72.2) | 48 (70.6) | 60 (75.9) | 52 (76.5) | 142 (74.3) | 176 (68.2) | 46 (68.7) | 136 (75.1) |
| Female | 153 (24.1) | 153 (28.0) | 22 (27.8) | 20 (29.4) | 19 (24.1) | 16 (23.5) | 49 (25.7) | 82 (31.8) | 21 (31.3) | 45 (24.9) |
| Age (y) | 45.1 ± 10.3 | 43.2 ± 10.0 | 47.8 ± 10.2 | 45.1 ± 8.5 | 46.9 ± 10.0 | 42.7 ± 9.2 | 50.6 ± 10.2 | 46.6 ± 10.4 | 44.7 ± 10.9 | 42.9 ± 9.8 |
| WHO pathological type | | | | | | | | | | |
| NKSCC | 613 (96.7) | 542 (99.1) | 78 (98.7) | 67 (98.5) | 72 (97.5) | 67 (98.5) | 169 (88.5) | 250 (96.9) | 66 (98.5) | 180 (99.4) |
| KSCC | 21 (3.3) | 5 (0.9) | 1 (1.3) | 1 (1.5) | 2 (2.5) | 1 (1.5) | 22 (11.5) | 8 (3.1) | 1 (1.5) | 1 (0.6) |
| Clinical T stage[a] | | | | | | | | | | |
| cT1 | 31 (4.9) | 131 (23.9) | 1 (1.3) | 16 (23.5) | 4 (5.1) | 17 (25.0) | 14 (7.3) | 41 (15.9) | 5 (7.4) | 14 (7.7) |
| cT2 | 113 (17.8) | 97 (17.7) | 11 (13.9) | 12 (17.6) | 14 (17.7) | 11 (16.2) | 23 (12.1) | 51 (19.8) | 6 (9.0) | 38 (21.0) |
| cT3 | 359 (56.6) | 273 (49.9) | 48 (60.8) | 35 (51.5) | 49 (62.0) | 29 (42.6) | 77 (40.3) | 120 (46.5) | 38 (56.7) | 102 (56.4) |
| cT4 | 131 (20.7) | 46 (8.4) | 19 (24.0) | 5 (7.4) | 12 (15.2) | 11 (16.2) | 77 (40.3) | 46 (17.8) | 18 (26.9) | 27 (14.9) |
| Clinical N stage[a] | | | | | | | | | | |
| cN0 | 31 (4.9) | 97 (17.7) | 3 (3.8) | 8 (11.8) | 3 (3.8) | 7 (10.3) | 15 (7.9) | 33 (12.8) | 4 (6.0) | 12 (6.6) |
| cN1 | 234 (36.9) | 253 (46.3) | 28 (35.4) | 33 (48.5) | 23 (29.1) | 35 (51.5) | 81 (42.4) | 107 (41.5) | 21 (31.3) | 69 (38.1) |
| cN2 | 275 (43.4) | 158 (28.9) | 36 (45.6) | 22 (32.4) | 38 (48.1) | 21 (30.9) | 74 (38.7) | 110 (42.6) | 27 (40.3) | 83 (45.9) |
| cN3 | 94 (14.8) | 39 (7.1) | 12 (15.2) | 5 (7.3) | 15 (19.0) | 5 (7.3) | 21 (11.0) | 8 (3.1) | 15 (22.4) | 17 (9.4) |
| Clinical stage[a] | | | | | | | | | | |
| I | 4 (0.6) | 37 (6.8) | 0 (0.0) | 3 (4.4) | 0 (0.0) | 3 (4.5) | 1 (0.5) | 7 (2.7) | 1 (1.5) | 2 (1.1) |
| II | 65 (10.3) | 115 (21.0) | 9 (11.4) | 15 (22.1) | 7 (8.8) | 16 (23.5) | 16 (8.4) | 45 (17.4) | 3 (4.5) | 25 (13.8) |
| III | 358 (56.5) | 312 (57.0) | 40 (50.6) | 41 (60.3) | 48 (60.8) | 33 (48.5) | 80 (41.9) | 154 (59.7) | 33 (49.3) | 113 (62.4) |
| IV | 207 (32.6) | 83 (15.2) | 30 (38.0) | 9 (13.2) | 24 (30.4) | 16 (23.5) | 94 (49.2) | 52 (20.2) | 30 (44.8) | 41 (22.7) |
| Pre-EBV DNA (copies/ml) | | | | | | | | | | |
| Available | 41,370 ± 214,841 | 42,959 ± 214,221 | 12,034 ± 31,208 | 16,484 ± 41,815 | 66,684 ± 230,918 | 33,318 ± 121,711 | NA | 7,601 ± 51,914 | 1,785 ± 4,204 | 12,700 ± 42,962 |
| NA | 136 (21.5) | 7 (1.3) | 10 (12.7) | 2 (2.9) | 9 (11.4) | 0 (0.0) | 191 (100.0) | 185 (71.7) | 1 (1.5) | 4 (2.2) |
| Chemotherapy | | | | | | | | | | |
| Yes | 556 (87.7) | 452 (82.6) | 72 (91.1) | 61 (89.7) | 74 (93.7) | 60 (88.2) | 188 (98.4) | 253 (98.1) | 62 (92.6) | 161 (89.0) |
| No | 78 (12.3) | 95 (17.4) | 7 (8.9) | 7 (10.3) | 5 (6.3) | 8 (11.8) | 3 (1.6) | 5 (1.9) | 5 (7.4) | 20 (11.0) |
| Radiotherapy technique | | | | | | | | | | |
| IMRT | 563 (88.8) | 546 (99.8) | 71 (89.9) | 68 (100.0) | 69 (87.3) | 68 (100.0) | 169 (88.5) | 258 (100.0) | 67 (100.0) | 181 (100.0) |
| 3D-CRT | 2 (0.3) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| 2D-CRT | 69 (10.9) | 1 (0.2) | 8 (10.1) | 0 (0.0) | 10 (12.7) | 0 (0.0) | 22 (11.5) | 0 (0.0) | 0 (0.0) | 0 (0.0) |
| Dose to $PTV_{nx}$ (cGy) | 6,972 ± 120 | 6,962 ± 101 | 6,985 ± 143 | 6,954 ± 96 | 6,977 ± 115 | 6,953 ± 89 | 6,879 ± 277 | 6,984 ± 104 | 6,995 ± 58 | 6,977 ± 96 |
| Dose to $PTV_{nd}$ (cGy) | 6,102 ± 1,693 | 5,994 ± 1,792 | 6,102 ± 1,629 | 6,173 ± 1,578 | 6,224 ± 1,496 | 6,099 ± 1,747 | 6,261 ± 977 | 6,380 ± 955 | 6,666 ± 858 | 6,435 ± 1,220 |

(*Continued on next page*)

**Table 1. *Continued***

| | Training (n = 1,181) | | Tuning (n = 147) | | Internal validation (n = 147) | | External validation (n = 449) | | Reader study (n = 248) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | rNPC (n = 634) | non-rNPC (n = 547) | rNPC (n = 79) | non-rNPC (n = 68) | rNPC (n = 79) | non-rNPC (n = 68) | rNPC (n = 191) | non-rNPC (n = 258) | rNPC (n = 67) | non-rNPC (n = 181) |
| Post-EBV DNA (copies/ml) | | | | | | | | | | |
| Available | 224.7 ± 2647.0 | 1.5 ± 25.1 | 1.2 ± 8.1 | 1.9 ± 13.7 | 665.3 ± 2701.0 | 41.2 ± 253.4 | 28.1 ± 95.5 | 3.7 ± 15.7 | 6.7 ± 51.3 | 13.5 ± 112.2 |
| NA | 191 (30.1) | 124 (22.7) | 30 (38.0) | 17 (25.0) | 26 (32.9) | 14 (20.6) | 145 (75.9) | 221 (85.7) | 2 (3.0) | 28 (15.5) |
| Follow-up interval (month) | 6.0 ± 3.4 | 8.5 ± 4.7 | 6.0 ± 3.6 | 8.4 ± 4.8 | 6.2 ± 3.9 | 8.4 ± 4.4 | 6.2 ± 4.6 | 6.6 ± 3.2 | 5.9 ± 3.3 | 7.8 ± 3.8 |

Data are n (%) or mean ± standard deviation.

Abbreviations: rNPC, recurrent nasopharyngeal carcinoma; non-rNPC, non-recurrent nasopharyngeal carcinoma; WHO, World Health Organization; NKSCC, non-keratinizing squamous cell carcinoma; KSCC, keratinizing squamous cell carcinoma; EBV DNA, Epstein-Barr virus deoxyribonucleic acid; Pre-EBV DNA, pretreatment plasma EBV DNA level; NA, not applicable; IMRT, intensity-modulated radiation therapy; 3D-CRT, three-dimensional conformal radiation therapy; 2D-CRT, two-dimensional conventional radiation therapy; PTVnx, primary tumor volume of nasopharynx; PTVnd, primary tumor volume of metastatic neck lymph node(s); Post-EBV DNA, posttreatment (3–6 months after radiation therapy) plasma EBV DNA level.

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

The RAIN detector preselected 10.4% positives and eliminated 86.8% negatives. Accordingly, the RAINMAN workflow resulted in a workload reduction of 83.6% scans. The ROC-AUC increased by 0.011–0.012, and the sensitivity increased by 1.5%–3.2%, while the specificity increased by 0.3%–2.9% among readers (Figure 5A and Table 2). Judgments per radiologist and the simulated workflow are presented in confusion matrices in Figure S2. There was a moderate degree of agreement among all combination pairs of readers with or without RAIN (Figure 5B).

## DISCUSSION

We developed and validated a triage-driven, semiautomated workflow with a good balance between the strengths and limitations of a fully automated analysis and manual reading. In a simulation study among large populations across varying recurrence rates 3.8%–10.2%, RAINMAN was competent in freeing radiologists from examining 81.3%–88.3% true-negative scans and in preselecting 10.4%–34.2% high-confident positives for enhanced assessment. Also, a boost in diagnostic performance was observed (sensitivity 1.5%–3.2%; specificity 0.3%–2.9%). Notably, this diagnosis refinement process motivated by human-computer cooperation can be extrapolated to any validated DL algorithm, any cancer and even noncancerous disease imaging, and any medical institution.

Although good harvests have been reaped in the field of CAD,[9,15,25–27] there will likely always be a need for a "human-in-the-loop" when complicated situations, especially those requiring an absolutely correct judgment (e.g., diagnosis in cancer or recurrence), are encountered.[28–30] Our simulated triage workflow, which builds on a convolutional neural network (CNN)-based tool and triage-driven approach, has several potential implications for managing cancer patients. Foremost, the uniqueness of RAIN in profiling risk probability longitudinally and dynamically conforms to the clinical setting. With simultaneously comprehensive iterating cues from both the present scan and prior scan(s), our model is able to eliminate the interference of histological and structural changes induced by radiation. Additionally, despite the diverse scanning protocols and scanner vendors between or within centers, the generalization and robustness of RAIN have been validated in a multicenter observational study. Inspired by promising and recent advancements in human-computer interactions,[16–18,29,30] we used the underlying risk prediction score generated by the DL algorithm for more prudent judgment. By deliberately leveraging the previous well-verified triage-driven approach,[29] this DL-based triage and radiologists-made decision assistance is more interpretable and transparent to the current clinical situation in reducing workloads while matching the diagnostic performance of radiologists.

Second, the high cure rate and certain recurrence rate of the studied cancer type (mostly recurrence free) in this work accord with resource-constrained scenarios.[4–6] Most routine follow-up MR scans among these patients with complete remission (CR) of NPC show stable histological and structural changes induced by radiation that may not require manual review. Thus, approaches to shunt these normal scans to reduce radiologist workload are needed. The automated shunt is able to dismiss a large portion (76.2%–83.6%) of the follow-up scans as recurrence free for no radiologist assessment which significantly improves workflow specificity and efficiency. Notably, this first-sift prescreening would cause a larger workload reduction in a realistic population with a lower recurrence rate than retrospectively collected populations. This result implies that the clinical utility of semiautomated workflow hinges on the cohort composition. With clinical resources becoming increasingly scarce with accumulating imaging data (mostly normal) and a relative decrease in experienced radiologists, our proposed workflow could sift out safe patients, facilitating more reasonable medical resource allocation. Obviating single-human

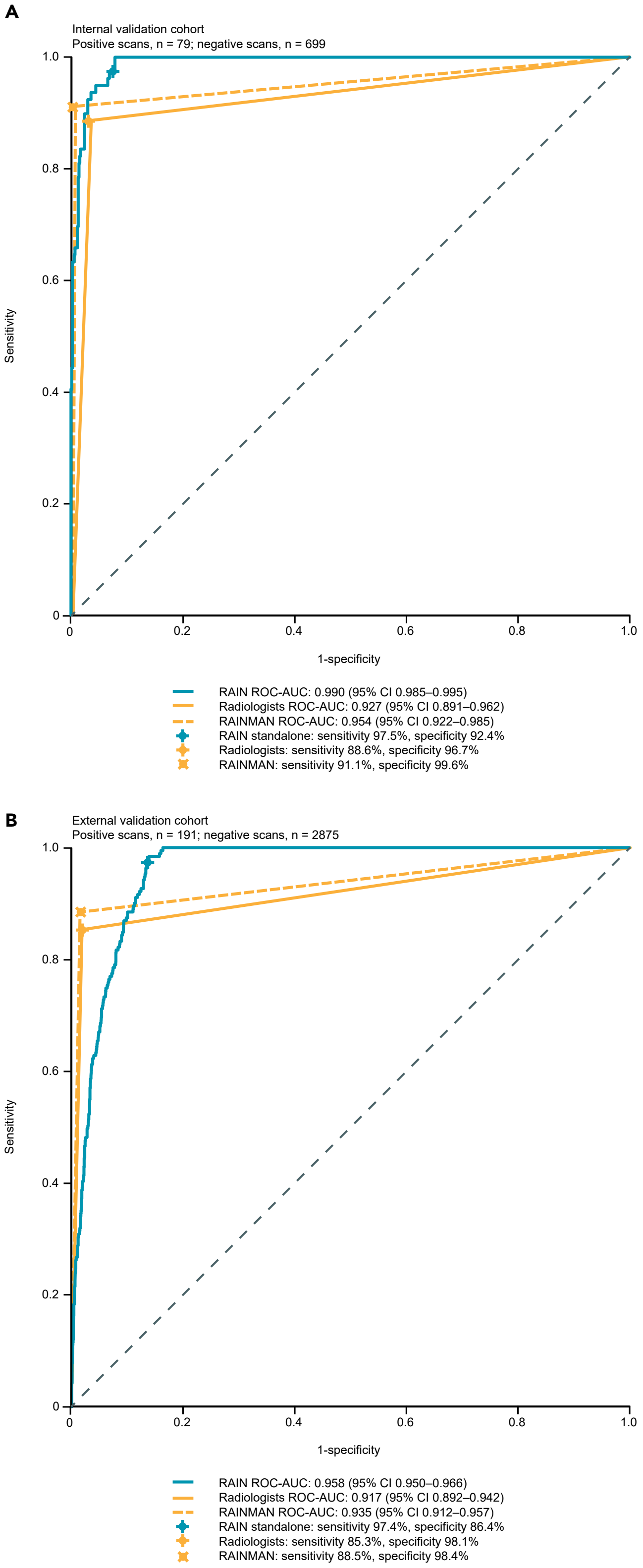

A
Internal validation cohort
Positive scans, n = 79; negative scans, n = 699
Sensitivity
1.0
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1.0
1-specificity
RAIN ROC-AUC: 0.990 (95% CI 0.985–0.995)
Radiologists ROC-AUC: 0.927 (95% CI 0.891–0.962)
RAINMAN ROC-AUC: 0.954 (95% CI 0.922–0.985)
RAIN standalone: sensitivity 97.5%, specificity 92.4%
Radiologists: sensitivity 88.6%, specificity 96.7%
RAINMAN: sensitivity 91.1%, specificity 99.6%
B
External validation cohort
Positive scans, n = 191; negative scans, n = 2875
Sensitivity
1.0
0.8
0.6
0.4
0.2
0
0
0.2
0.4
0.6
0.8
1.0
1-specificity
RAIN ROC-AUC: 0.958 (95% CI 0.950–0.966)
Radiologists ROC-AUC: 0.917 (95% CI 0.892–0.942)
RAINMAN ROC-AUC: 0.935 (95% CI 0.912–0.957)
RAIN standalone: sensitivity 97.4%, specificity 86.4%
Radiologists: sensitivity 85.3%, specificity 98.1%
RAINMAN: sensitivity 88.5%, specificity 98.4%

**Figure 3. Performances of RAINMAN and original radiologists on retrospective validation cohorts**

(A ) Performances of RAINMAN and original radiologists on internal validation cohort.

(B) Performances of RAINMAN and original radiologists on external validation cohort.

Abbreviations: ROC-AUC, area under curve the receiver operating characteristic curve; RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; RAINMAN, RAIN + MANual reading. See also Table S3.

assessment in normal scans (both time and discussion) potentially allows radiologists to focus on complicated and atypical scans that need human interpretations.

Last but surprisingly, we found that the DL-motivated examination triage was able to preselect a subset with high risk. Accordingly, the sensitivity in detecting recurrences was markedly improved by complementary human-computer cooperation. Undeniably, owing to histological and structural changes induced by radiation, disturbance of subjective judgment, and heavy workload, clinicians may inevitably neglect some occult and atypical recurrence or mistake them for benign "mimickers."[10,11] Thus, a few individuals might end up with recurrence that, in hindsight, could have been detected on previous scans.[31] The impartial judgment provided by the DL method may potentially reduce omissions in recurrence wavered by subjective influences. If the proposed workflow is implemented in clinical scenarios, there would be not only a reduction in the number of scan-detected recurrences but also an expected shift toward downstaging recurrence in the future.

In conclusion, our study shows that leveraging risk scores predicted by DL model and existing heuristics of radiologist decision-making in a triage-based approach that triages scans into mutually exclusive tiers could potentially reduce radiologist workload by over half and impartially present positives with high confidence. The described triage-driven approach offers a logical mode for status-wise clinical adoption and performance testing in management among cancer patients.

## Limitations of the study

Several limitations should be noted in our work. A limitation of our study was that, while our model was built and evaluated on diverse and clinically representative cases from three expert cancer-care centers (diverse follow-up intervals, imaging protocols, and scanner vendors), the subjects were all from China.[3] Additional cohorts with diversities in populations, races, and recurrence rates are required to confirm its

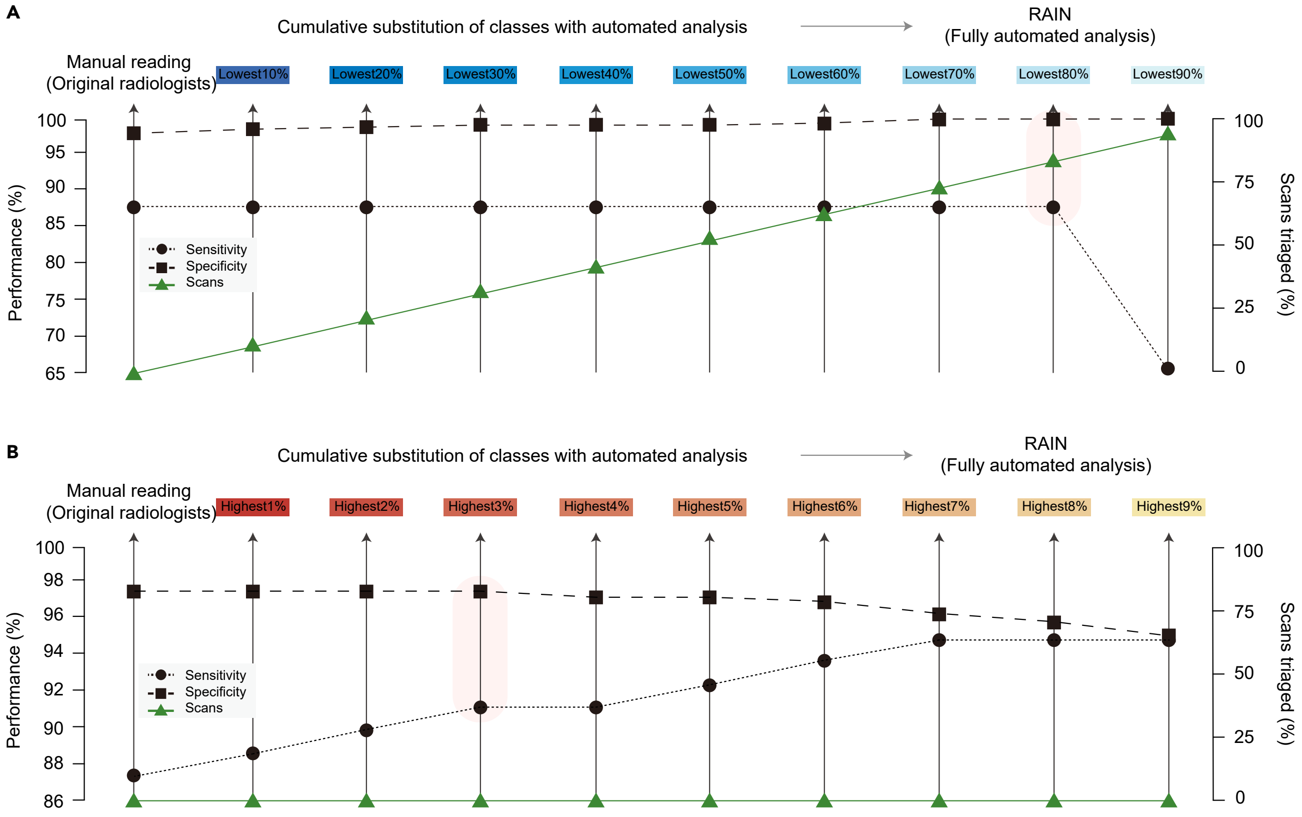


**Figure 4. Generation of the rule-out and alert thresholds through triage substitution scheme on tuning cohort**

The red rectangle in (A) indicates a stage retaining the undamaged sensitivity with maximum workload substitution. The red rectangle in (B) indicates a stage retaining undamaged specificity. Abbreviation: RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma. See also Table S4.

Table 2. Diagnostic performance and workload reduction simulation for radiologist(s) and RAINMAN workflow in retrospective and reader study

| | RAIN (%) [95% CI] | Manual read (%) [95% CI] | RAINMAN (%) [95% CI] | Δ change (%) | p value |
|---|---|---|---|---|---|
| Retrospective study | | | | | |
| Internal validation cohort | | | | | |
| Original radiologists | | | | | |
| ROC-AUC | 0.990 (0.984–0.995) | 0.927 (0.891–0.962) | 0.954 (0.922–0.985) | +0.027 | 0.004 |
| Sensitivity | 97.5 (77/79) [93.9–100.0] | 88.6 (70/79) [81.4–95.8] | 91.1(72/79) [84.7–97.5] | +2.5 | 0.500 |
| Specificity | 92.4 (646/669) [90.5–94.4] | 96.7 (676/699) [95.4–98.0] | 99.6 (696/699) [99.1–100.0] | +2.9 | <0.001 |
| Recurrence alert | 97.5 (77/79) | 0.0 (0/79) | 34.2(27/79) | NA | NA |
| Non-recurrence elimination | 92.4 (646/669) | 0.0 (0/699) | 88.3(617/699) | NA | NA |
| Workload reduction | 100.0 (778/778) | 0.0 (0/778) | 79.3(617/778) | NA | NA |
| External validation cohort | | | | | |
| Original radiologists | | | | | |
| ROC-AUC | 0.958 (0.950–0.966) | 0.917 (0.892–0.942) | 0.935 (0.912–0.957) | +0.018 | 0.007 |
| Sensitivity | 97.4 (186/191) [95.1–99.7] | 85.3(163/191) [80.3–90.4] | 88.5 (169/191) [83.9–93.1] | +3.2 | 0.031 |
| Specificity | 86.4 (2,484/2,875) [85.1–87.6] | 98.1(2,820/2,875) [97.6–98.6] | 98.4 (2,830/2,875) [98.0–98.9] | +0.3 | 0.302 |
| Recurrence alert | 97.4 (186/191) | 0.0 (0/191) | 29.8 (57/191) | NA | NA |
| Non-recurrence elimination | 86.4 (2,484/2,875) | 0.0 (0/2,875) | 81.3(2,337/2,875) | NA | NA |
| Workload reduction | 100.0 (3,066/3,066) | 0.0 (0/3,066) | 76.2(2,337/3,066) | NA | NA |
| Reader study | | | | | |
| Radiologist 1 | | | | | |
| ROC-AUC | 0.969 (0.960–0.979) | 0.908 (0.863–0.953) | 0.923(0.880–0.966) | +0.015 | 0.053 |
| Sensitivity | 91.0 (61/67) [84.0–98.1] | 83.6 (56/67) [74.5–92.7] | 85.1(57/67) [76.3–93.8] | +1.5 | 1.000 |
| Specificity | 91.7 (1,529/1,667) [90.4–93.0] | 98.1(1,676/1,709) [97.4–98.7] | 99.4 (1,701/1,709) [99.2–99.9] | +1.3 | <0.001 |
| Recurrence alert | 91.0 (61/67) | 0.0 (0/67) | 10.4 (7/67) | NA | NA |
| Non-recurrence elimination | 91.7 (1,529/1,667) | 0.0 (0/1,709) | 86.8 (1,484/1,709) | NA | NA |
| Workload reduction | 100.0 (1,776/1,776) | 0.0 (0/1,776) | 83.6 (1,484/1,776) | NA | NA |
| Radiologist 2 | | | | | |
| ROC-AUC | 0.969 (0.960–0.979) | 0.934 (0.895–0.974) | 0.946 (0.909–0.983) | +0.012 | 0.137 |
| Sensitivity | 91.0 (61/67) [84.0–98.1] | 88.1(59/67) [80.1–96.0] | 89.6 (60/67) [82.0–97.1] | +1.5 | 1.000 |
| Specificity | 91.7 (1,529/1,667) [90.4–93.0] | 98.8 (1,689/1,709) [98.3–99.3] | 99.6 (1,702/1,709) [99.3–99.9] | +0.8 | 0.007 |
| Recurrence alert | 91.0 (61/67) | 0.0 (0/67) | 10.4 (7/67) | NA | NA |
| Non-recurrence elimination | 91.7 (1,529/1,667) | 0.0 (0/1,709) | 86.8 (1,484/1,709) | NA | NA |
| Workload reduction | 100.0 (1,776/1,776) | 0.0 (0/1,776) | 83.6 (1,484/1,776) | NA | NA |
| Radiologist 3 | | | | | |
| ROC-AUC | 0.969 (0.960–0.979) | 0.950 (0.916–0.985) | 0.961(0.929–0.992) | +0.011 | 0.170 |
| Sensitivity | 91.0 (61/67) [84.0–98.1] | 91.0 (61/67) [84.0–98.1] | 92.5 (62/67) [86.1–99.0] | +1.5 | 1.000 |
| Specificity | 91.7 (1,529/1,667) [90.4–93.0] | 99.0 (1,692/1,709) [98.5–99.5] | 99.6 (1,702/1,709) [99.3–99.9] | +0.6 | 0.031 |
| Recurrence alert | 91.0 (61/67) | 0.0 (0/67) | 10.4 (7/67) | NA | NA |
| Non-recurrence elimination | 91.7 (1,529/1,667) | 0.0 (0/1,709) | 86.8 (1,484/1,709) | NA | NA |
| Workload reduction | 100.0 (1,776/1,776) | 0.0 (0/1,776) | 83.6 (1,484/1,776) | NA | NA |
| Mean radiologist | | | | | |
| ROC-AUC | 0.969 (0.960–0.979) | 0.931(0.891–0.971) | 0.943(0.906–0.980) | +0.012 | – |

(*Continued on next page*)

**Table 2.** ***Continued***

| | RAIN (%) [95% CI] | Manual read (%) [95% CI] | RAINMAN (%) [95% CI] | Δ change (%) | p value |
|---|---|---|---|---|---|
| Sensitivity | 91.0 (61/67) [84.0–98.1] | 87.6 (59/67) [79.5–95.6] | 89.1(60/67) [81.5–96.6] | +1.5 | 1.000 |
| Specificity | 91.7 (1,529/1,667) [90.4–93.0] | 98.6 (1,685/1,709) [97.7–99.2] | 99.5 (1,700/1,709) [99.3–99.9] | +0.9 | <0.001 |
| Recurrence alert | 91.0 (61/67) | 0.0 (0/67) | 10.4 (7/67) | NA | NA |
| Non-recurrence elimination | 91.7 (1,529/1,667) | 0.0 (0/1,709) | 86.8 (1,484/1,709) | NA | NA |
| Workload reduction | 100.0 (1,776/1,776) | 0.0 (0/1,776) | 83.6 (1,484/1,776) | NA | NA |

Data are number with 95% confidential interval (CI) in parentheses, or percentages with raw data in parentheses and 95% CI in brackets.
Abbreviations: RAINMAN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma + MANual reading; CI, confidential interval.
The operating points adopted in the RAINMAN workflow are based on two thresholds that triaged scans into three tiers, including high-confidence negatives for automated read, high-confidence positives, and equivocal cases for manual reading. The threshold setting and selection of operating points on the tuning cohort are presented in the Figure 4. Δ indicates differences in sensitivity, specificity, and workload reduction when the RAINMAN workflow is simulated. p values were calculated using the Delong's test or McNemar test.

performance. Another limitation was that cases of partial remission (PR) of the tumor (3.0%–13.0%)[4–6] were not considered in this study. Third, because of the unreachable sites of some recurring disease, not all cases were diagnosed with histology. However, we examined nonpathologically confirmed recurrences according to strict criteria. Besides, certain occult recurrence may occur before clinical diagnostic recurrence. The date of recurrence is unlikely to correspond exactly to when the scan was acquired. A final limitation was that the thresholds for the DL algorithm were derived in our setting, and the results from other centers could be different.

## STAR★METHODS

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

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - ○ Patient cohorts
- METHOD DETAILS
  - ○ Data collection
  - ○ Therapeutic regimens and follow-up
  - ○ Clinical classification
  - ○ MR scanning protocol
  - ○ MR scan labeling
  - ○ MR registration and network architectures
  - ○ Establishment of triage-driven semiautomated workflow
  - ○ Reader study
- QUANTIFICATION AND STATISTICAL ANALYSIS
- ADDITIONAL RESOURCES

## SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.isci.2023.108347.

## ACKNOWLEDGMENTS

This study was supported by grants from the National Natural Science Foundation of China (No. 82172862 and No. 82172863) and the Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515010118). The paper has been edited by AJE Editing Services (verification code: 80F0-657F-D66D-EB79-8C3P). Figure 1 was created with BioRender.com with an academic license.

## AUTHOR CONTRIBUTIONS

Conceptualization, C.-F.L., X.L., L.-R.K., and X.G.; methodology, Y.-Y.H., Y.-S.D., Y.L., M.-Y.Q., W.-Z.Q., B.-Z.J., C.-Y.F., H.-H.C., X.C., H.-Y.H., Y.S., and C.-F.L.; software, Y.-Y.H., Y.-S.D., B.-Z.J., C.-Y.F., H.-H.C., X.C., H.-Y.H., and C.-F.L.; validation, Y.-Y.H., Y.-S.D., Y.L., M.-Y.Q., and W.-Z.Q.; formal analysis, Y.-Y.H., Y.-S.D., Y.L., M.-Y.Q., W.-Z.Q., B.-Z.J., C.-Y.F., H.-H.C., X.C., J.-Y.Z., H.-Y.H., Z.-J.Z., and Y.D.; investigation,

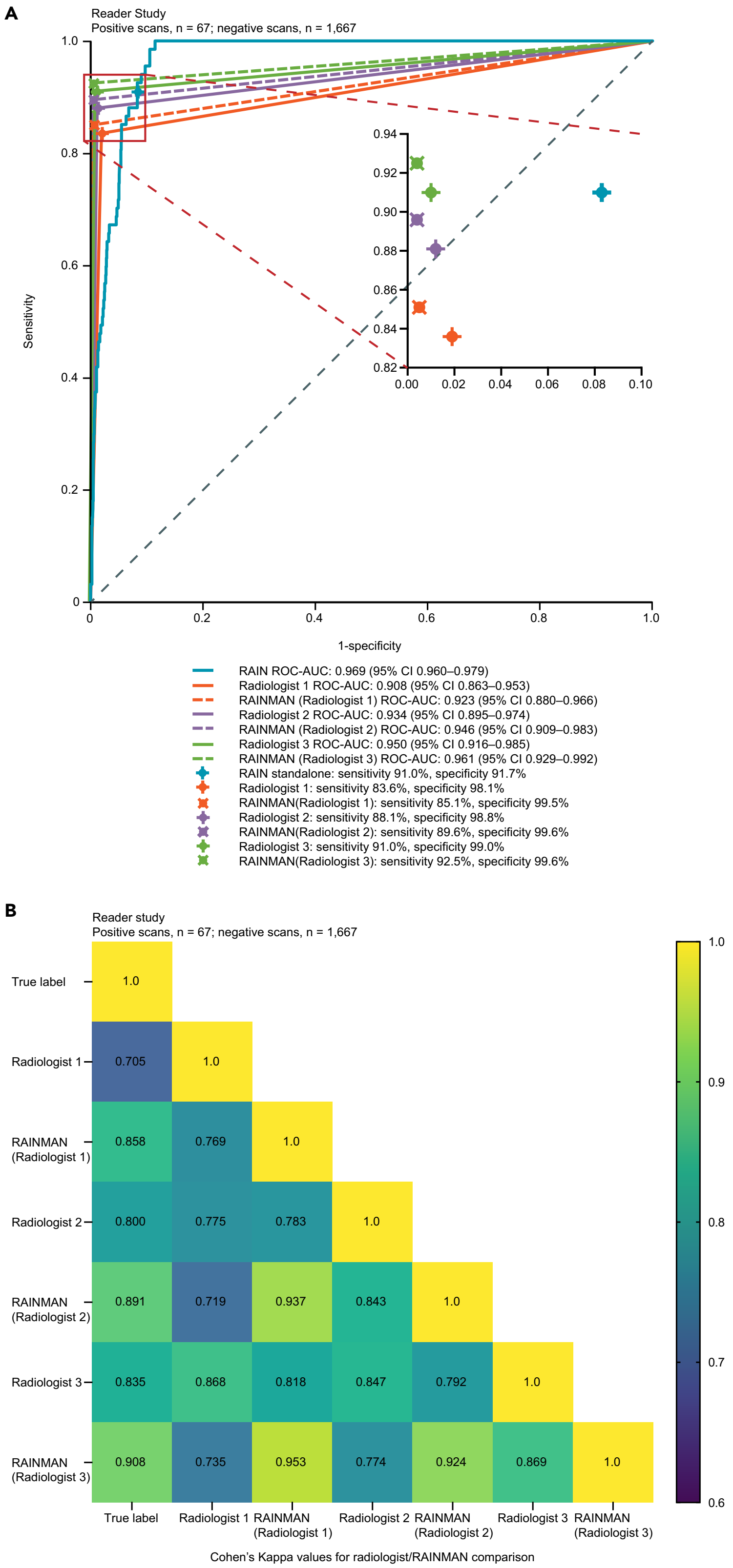

A
Reader Study
Positive scans, n = 67; negative scans, n = 1,667
Sensitivity
1-specificity
0
0.2
0.4
0.6
0.8
1.0
0.82
0.84
0.86
0.88
0.90
0.92
0.94
0.00
0.02
0.04
0.06
0.08
0.10
RAIN ROC-AUC: 0.969 (95% CI 0.960–0.979)
Radiologist 1 ROC-AUC: 0.908 (95% CI 0.863–0.953)
RAINMAN (Radiologist 1) ROC-AUC: 0.923 (95% CI 0.880–0.966)
Radiologist 2 ROC-AUC: 0.934 (95% CI 0.895–0.974)
RAINMAN (Radiologist 2) ROC-AUC: 0.946 (95% CI 0.909–0.983)
Radiologist 3 ROC-AUC: 0.950 (95% CI 0.916–0.985)
RAINMAN (Radiologist 3) ROC-AUC: 0.961 (95% CI 0.929–0.992)
RAIN standalone: sensitivity 91.0%, specificity 91.7%
Radiologist 1: sensitivity 83.6%, specificity 98.1%
RAINMAN(Radiologist 1): sensitivity 85.1%, specificity 99.5%
Radiologist 2: sensitivity 88.1%, specificity 98.8%
RAINMAN(Radiologist 2): sensitivity 89.6%, specificity 99.6%
Radiologist 3: sensitivity 91.0%, specificity 99.0%
RAINMAN(Radiologist 3): sensitivity 92.5%, specificity 99.6%
B
Reader study
Positive scans, n = 67; negative scans, n = 1,667
True label 1.0
Radiologist 1 0.705 1.0
RAINMAN (Radiologist 1) 0.858 0.769 1.0
Radiologist 2 0.800 0.775 0.783 1.0
RAINMAN (Radiologist 2) 0.891 0.719 0.937 0.843 1.0
Radiologist 3 0.835 0.868 0.818 0.847 0.792 1.0
RAINMAN (Radiologist 3) 0.908 0.735 0.953 0.774 0.924 0.869 1.0
True label
Radiologist 1
RAINMAN (Radiologist 1)
Radiologist 2
RAINMAN (Radiologist 2)
Radiologist 3
RAINMAN (Radiologist 3)
Cohen's Kappa values for radiologist/RAINMAN comparison
1.0
0.9
0.8
0.7
0.6

**Figure 5. Performances of RAIN, readers, and simulated RAINMAN, and the degree of agreement for readers and simulated RAINMAN based on reader study**

(A) Performances of RAIN, readers, and simulated RAINMAN.

(B) Degree of agreement for readers and simulated RAINMAN. Abbreviations: RAIN, Artificial Intelligence for detecting Recurrent Nasopharyngeal carcinoma; ROC-AUC, area under curve the receiver operating characteristic curve; RAINMAN, RAIN + MANual reading. See also Figure S2 and Table S3.

Y.-Y.H., Y.-S.D., Y.L., M.-Y.Q., W.-Z.Q., W.-X.X., B.-Z.J.; resources, C.-F.L., X.L., L.-R.K., X.G., C.-M.X., Y.S., H.-Q.M., and L.-Q.T.; data curation, Y.-Y.H., Y.-S.D., Y.L., M.-Y.Q., W.-Z.Q., C.-Y.F., H.-H.C., J.-Y.Z., H.-Y.H., Z.-J.Z., and Y.D.; writing – original draft, Y.-Y.H., Y.-S.D., Y.L., M.-Y.Q., W.-Z.Q., and B.-Z.J.; writing-review & editing, C.-F.L., X.L., L.-R.K., and X.G.; visualization, Y.-Y.H., Y.-S.D., Y.L., M.-Y.Q., and W.-Z.Q.; supervision, C.-F.L., X.L., L.-R.K., X.G., C.-M.X., Y.S., H.-Q.M., and L.-Q.T.; project administration, C.-F.L., X.L., L.-R.K., X.G., C.-M.X., Y.S., H.-Q.M., and L.-Q.T.; funding acquisition, C.-F.L., and X.L.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

## INCLUSION AND DIVERSITY

We support inclusive, diverse, and equitable conduct of research.

Received: June 27, 2023
Revised: August 8, 2023
Accepted: October 24, 2023
Published: October 29, 2023

## REFERENCES

1. Shapiro, C.L. (2018). Cancer Survivorship. N. Engl. J. Med. *379*, 2438–2450.
2. Mahvi, D.A., Liu, R., Grinstaff, M.W., Colson, Y.L., and Raut, C.P. (2018). Local Cancer Recurrence: The Realities, Challenges, and Opportunities for New Therapies. CA A Cancer J. Clin. *68*, 488–505.
3. Siegel, R.L., Miller, K.D., Fuchs, H.E., and Jemal, A. (2021). Cancer Statistics, 2021. CA. Cancer J. Clin. *71*, 7–33.
4. Lee, A.W.M., Ng, W.T., Chan, L.L.K., Hung, W.M., Chan, C.C.C., Sze, H.C.K., Chan, O.S.H., Chang, A.T.Y., and Yeung, R.M.W. (2014). Evolution of treatment for nasopharyngeal cancer – Success and setback in the intensity-modulated radiotherapy era. Radiother. Oncol. *110*, 377–384.
5. Zhang, M.-X., Li, J., Shen, G.-P., Zou, X., Xu, J.-J., Jiang, R., You, R., Hua, Y.-J., Sun, Y., Ma, J., et al. (2015). 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.
6. Mao, Y.-P., Tang, L.-L., Chen, L., Sun, Y., Qi, Z.-Y., Zhou, G.-Q., Liu, L.-Z., Li, L., Lin, A.-H., and Ma, J. (2016). Prognostic factors and failure patterns in non-metastatic nasopharyngeal carcinoma after intensity-modulated radiotherapy. Chin. J. Cancer *35*, 103.
7. Pfister, D.G., Spencer, S., Adelstein, D., Adkins, D., Anzai, Y., Brizel, D.M., Bruce, J.Y., Busse, P.M., Caudell, J.J., Cmelak, A.J., et al. (2020). Head and Neck Cancers, Version 2.2020, NCCN Clinical Practice Guidelines in Oncology. J. Natl. Compr. Cancer Netw. *18*, 873–898.
8. Tang, L.-L., Chen, Y.-P., Chen, C.-B., Chen, M.-Y., Chen, N.-Y., Chen, X.-Z., Du, X.-J., Fang, W.-F., Feng, M., Gao, J., et al. (2021). The Chinese Society of Clinical Oncology (CSCO) clinical guidelines for the diagnosis and treatment of nasopharyngeal carcinoma. Cancer Commun. *41*, 1195–1227.
9. Bi, W.L., Hosny, A., Schabath, M.B., Giger, M.L., Birkbak, N.J., Mehrtash, A., Allison, T., Arnaout, O., Abbosh, C., Dunn, I.F., et al. (2019). Artificial intelligence in cancer imaging: Clinical challenges and applications. CA. Cancer J. Clin. *69*, 127–157.
10. Meng, K., Tey, J., Ho, F.C.H., Asim, H., and Cheo, T. (2020). Utility of magnetic resonance imaging in determining treatment response and local recurrence in nasopharyngeal carcinoma treated curatively. BMC Cancer *20*, 193.
11. Fitzgerald, R. (2001). Error in radiology. Clin. Radiol. *56*, 938–946.
12. Hong, T.S., Tomé, W.A., and Harari, P.M. (2012). Heterogeneity in head and neck IMRT target design and clinical practice. Radiother. Oncol. *103*, 92–98.
13. Hwang, E.J., Lee, J.S., Lee, J.H., Lim, W.H., Kim, J.H., Choi, K.S., Choi, T.W., Kim, T.-H., Goo, J.M., and Park, C.M. (2021). Deep Learning for Detection of Pulmonary Metastasis on Chest Radiographs. Radiology *301*, 455–463.
14. Hiremath, A., Shiradkar, R., Fu, P., Mahran, A., Rastinehad, A.R., Tewari, A., Tirumani, S.H., Purysko, A., Ponsky, L., and Madabhushi, A. (2021). An integrated nomogram combining deep learning, Prostate Imaging-Reporting and Data System (PI-RADS) scoring, and clinical variables for identification of clinically significant prostate cancer on biparametric MRI: a retrospective multicentre study. Lancet. Digit. Health *3*, e445–e454.
15. Luo, H., Xu, G., Li, C., He, L., Luo, L., Wang, Z., Jing, B., Deng, Y., Jin, Y., Li, Y., et al. (2019). Real-time artificial intelligence for detection of upper gastrointestinal cancer by endoscopy: a multicentre, case-control, diagnostic study. Lancet Oncol. *20*, 1645–1654.
16. Yala, A., Schuster, T., Miles, R., Barzilay, R., and Lehman, C. (2019). A Deep Learning Model to Triage Screening Mammograms: A Simulation Study. Radiology *293*, 38–46.
17. Raya-Povedano, J.L., Romero-Martín, S., Elías-Cabot, E., Gubern-Mérida, A., Rodríguez-Ruiz, A., and Álvarez-Benito, M. (2021). AI-based Strategies to Reduce Workload in Breast Cancer Screening with Mammography and Tomosynthesis: A Retrospective Evaluation. Radiology *300*, 57–65.
18. Dembrower, K., Wåhlin, E., Liu, Y., Salim, M., Smith, K., Lindholm, P., Eklund, M., and Strand, F. (2020). Effect of artificial intelligence-based triaging of breast cancer screening mammograms on cancer detection and radiologist workload: a retrospective simulation study. Lancet. Digit. Health *2*, e468–e474.
19. Lång, K., Dustler, M., Dahlblom, V., Åkesson, A., Andersson, I., and Zackrisson, S. (2021). Identifying normal mammograms in a large screening population using artificial intelligence. Eur. Radiol. *31*, 1687–1692.
20. Leibig, C., Brehmer, M., Bunk, S., Byng, D., Pinker, K., and Umutlu, L. (2022). Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis. Lancet. Digit. Health *4*, e507–e519.
21. Shi, Y., Zu, C., Yang, P., Tan, S., Ren, H., Wu, X., Zhou, J., and Wang, Y. (2023). Uncertainty-weighted and relation-driven consistency training for semi-supervised head-and-neck tumor segmentation. Knowl. Base Syst. *272*, 110598.
22. Tang, P., Yang, P., Nie, D., Wu, X., Zhou, J., and Wang, Y. (2022). Unified medical image segmentation by learning from uncertainty in an end-to-end manner. Knowl. Base Syst. *241*, 108215.

23. Wong, L.M., King, A.D., Ai, Q.Y.H., Lam, W.K.J., Poon, D.M.C., Ma, B.B.Y., Chan, K.C.A., and Mo, F.K.F. (2021). Convolutional neural network for discriminating nasopharyngeal carcinoma and benign hyperplasia on MRI. Eur. Radiol. *31*, 3856–3863.

24. Ke, L., Deng, Y., Xia, W., Qiang, M., Chen, X., Liu, K., Jing, B., He, C., Xie, C., Guo, X., et al. (2020). Development of a self-constrained 3D DenseNet model in automatic detection and segmentation of nasopharyngeal carcinoma using magnetic resonance images. Oral Oncol. *110*, 104862.

25. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., and Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature *542*, 115–118.

26. Yun, T.J., Choi, J.W., Han, M., Jung, W.S., Choi, S.H., Yoo, R.-E., and Hwang, I.P. (2023). Deep learning based automatic detection algorithm for acute intracranial haemorrhage: a pivotal randomized clinical trial. NPJ Digit. Med. *6*, 61.

27. Kim, Y., Lee, K.J., Sunwoo, L., Choi, D., Nam, C.-M., Cho, J., Kim, J., Bae, Y.J., Yoo, R.-E., Choi, B.S., et al. (2019). Deep Learning in Diagnosis of Maxillary Sinusitis Using Conventional Radiography. Invest. Radiol. *54*, 7–15.

28. Labus, S., Altmann, M.M., Huisman, H., Tong, A., Penzkofer, T., Choi, M.H., Shabunin, I., Winkel, D.J., Xing, P., Szolar, D.H., et al. (2023). A concurrent, deep learning-based computer-aided detection system for prostate multiparametric MRI: a performance study involving experienced and less-experienced radiologists. Eur. Radiol. *33*, 64–76.

29. Gehrung, M., Crispin-Ortuzar, M., Berman, A.G., O'Donovan, M., Fitzgerald, R.C., and Markowetz, F. (2021). Triage-driven diagnosis of Barrett's esophagus for early detection of esophageal adenocarcinoma using deep learning. Nat. Med. *27*, 833–841.

30. Tschandl, P., Rinner, C., Apalla, Z., Argenziano, G., Codella, N., Halpern, A., Janda, M., Lallas, A., Longo, C., Malvehy, J., et al. (2020). Human-computer collaboration for skin cancer recognition. Nat. Med. *26*, 1229–1234.

31. Yala, A., Mikhael, P.G., Lehman, C., Lin, G., Strand, F., Wan, Y.-L., Hughes, K., Satuluru, S., Kim, T., Banerjee, I., et al. (2022). Optimizing risk-based breast cancer screening policies with reinforcement learning. Nat. Med. *28*, 136–143.

32. Amin, M.B., Greene, F.L., Edge, S.B., Compton, C.C., Gershenwald, J.E., Brookland, R.K., Meyer, L., Gress, D.M., Byrd, D.R., and Winchester, D.P. (2017). The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging. CA A Cancer J. Clin. *67*, 93–99.

33. Therasse, P., Arbuck, S.G., Eisenhauer, E.A., Wanders, J., Kaplan, R.S., Rubinstein, L., Verweij, J., Van Glabbeke, M., van Oosterom, A.T., Christian, M.C., and Gwyther, S.G. (2000). New Guidelines to Evaluate the Response to Treatment in Solid Tumors. J. Natl. Cancer Inst. *92*, 205–216.

34. Ng, S.-H., Chang, J.T.-C., Chan, S.-C., Ko, S.-F., Wang, H.-M., Liao, C.-T., Chang, Y.-C., and Yen, T.-C. (2004). Nodal metastases of nasopharyngeal carcinoma: patterns of disease on MRI and FDG PET. Eur. J. Nucl. Med. Mol. Imag. *31*, 1073–1080.

35. Ng, S.-H., Chan, S.-C., Yen, T.-C., Liao, C.-T., Chang, J.T.-C., Ko, S.-F., Wang, H.-M., Lin, C.-Y., Chang, K.-P., and Lin, Y.-C. (2010). Comprehensive imaging of residual/recurrent nasopharyngeal carcinoma using whole-body MRI at 3 T compared with FDG-PET-CT. Eur. Radiol. *20*, 2229–2240.

36. Chong, V.F., and Fan, Y.F. (1997). Detection of recurrent nasopharyngeal carcinoma: MR imaging versus CT. Radiology *202*, 463–470.

37. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., and Gee, J.C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. Neuroimage *54*, 2033–2044.

38. He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep Residual Learning for Image Recognition. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (IEEE), pp. 770–778.

39. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. (2021). An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale.

40. Ben-Baruch, E., Ridnik, T., Zamir, N., Noy, A., Friedman, I., Protter, M., and Zelnik-Manor, L. (2021). Asymmetric Loss for Multi-Label Classification.

41. Pérez-García, F., Sparks, R., and Ourselin, S. (2021). TorchIO: A Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. Comput. Methods Programs Biomed. *208*, 106236.

# STAR★METHODS

## KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---|---|---|
| Software and algorithms | | |
| Horos DICOM VIEWER (version 3.3.6) | Horos | https://www.horosproject.org |
| RadiAnt DICOM VIEWER (version 5.0.1) | RadiAnt | https://www.radiantviewer.cn |
| Python (version 3.6.10) | Python software | www.python.org |
| ANTsPy | ANTsPy | https://antspyx.readthedocs.io/en/latest/registration.html |
| ResNet18 | (He et al., 2016) | https://openaccess.thecvf.com/content_cvpr_2016/papers/He_Deep_Residual_Learning_CVPR_2016_paper.pdf |
| Transformer | (Dosovitskiy et al., 2021) | https://openreview.net/pdf?id=YicbFdNTTy |
| R (version 3.5.1) | R software | https://www.r-project.org |
| SPSS (version 26.0.0.0) | IBM corporation | https://www.ibm.com/analytics |
| GraphPad Prism (version 9.0.2) | GraphPad Software | https://www.graphpad.com |
| Other | | |
| Research Data Deposit | Sun Yat-sen University Cancer Center | https://www.researchdata.org.cn |
| Source code | Github | https://github.com/yydashu/RAIN |
| AJCC TNM staging system | (Amin et al., 2017) | https://link.springer.com/book/9783319406176 |

## RESOURCE AVAILABILITY

### Lead contact

Further information about the methods and requests for data or scripts should be directed to and will be fulfilled by the lead contact, Chao-Feng Li (lichaofeng@sysucc.org.cn).

### Materials availability

This study did not generate new unique reagents.

### Data and code availability

- De-identified patient standardized data have been deposited at the Research Data Deposit public platform (No. RDDA2023643426), and DOIs are listed in the key resources table. They are available upon request if access is granted. To request access, contact Sun Yat-sen University Cancer Center.
- All original code has been deposited at the Github and is publicly available as of the date of publication. DOIs are listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request (lichaofeng@sysucc.org.cn).

## EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

### Patient cohorts

In this multicentre retrospective study, patients with CR of non-disseminated NPC after radical (chemo)radiotherapy and followed annual H&N MR scans within 5 years between September 2007 and December 2021 were recruited from three hospitals in China and followed up until June 31, 2022.

Patients who met the following inclusion criteria were enrolled in this study: (i) aged ≥18 years with pathologically confirmed NPC at the initial diagnosis; (ii) restaged to I-IVa ($T_{1\text{-}4}N_{0\text{-}3}M_0$) according to the American Joint Committee on Cancer (AJCC) 8th edition staging classification system[32]; (iii) treated with radical radiotherapy (RT) ± neoadjuvant or concurrent chemotherapy; (iv) imaging confirmation of CR of gross tumor at the primary site and metastatic lymph nodes in the cervical region within 6 months post RT[33]; (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. 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 of RT > 7 days; (iii) absence or had pretreatment or follow-up MR images of insufficient quality to obtain measurements (i.e., motion artifacts); (iv) unknown efficacy evaluation or partial remission (PR) or stable disease (SD) beyond 6 months after RT[33]; (v) irregular follow-up surveys; or (vi) other malignancies or previous antitumour therapy.

Between September 10, 2007, and April 14, 2020, 792 eligible patients (1092 male) with CR of NPC who underwent annual follow-up MR surveys were collected from the Sun Yat-sen University Cancer Center (SYSUCC) and served as the training, tuning and internal validation cohorts. For external validation, 364 eligible patients (261 male) between July 20, 2009, and November 24, 2020, from Cancer Hospital of The University of Chinese Academy of Sciences (CHUCAS), and 85 eligible patients (57 male) between September 6, 2011, and September 4, 2020, from The Affiliated Cancer Hospital of Guangzhou Medical University (ACHGMU) were collected. From January to December 2021, 248 eligible cases (182 male) in SYSUCC were consecutively collected for comparing performances of RAIN model and three radiologists in the reader study. The institutional review boards of each center approved this study and waived the requirement for informed consent owing to the retrospective nature of this study.

## METHOD DETAILS

### Data collection

A semiautomated workflow (RAINMAN) was constructed post hoc based on a retrospective, observational study (ChiCTR.org.cn, ChiCTR2200056595) in three centers (SYSUCC; CHUCAS; ACHGMU). The study flowchart is presented in Figure 1.

### Therapeutic regimens and follow-up

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

After treatment, patients were assessed every 3 months during the first 3 years, every 6 months during the next 2 years, and annually thereafter. The routine follow-up H&N MR scans were acquired at regular intervals within 5 years, generally at 3–24 months, depending on the oncologists' preference and survivors' compliance.

### Clinical classification

A rotating panel of four experts (two radiologists and two oncologists) specialized in NPC evaluated patients' clinical information and all scans. 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. CR was defined as absence of any disease, including both the primary tumor at the nasopharynx and the metastatic cervical lymph node(s), on MR images.

During 5-year follow-up period, patients were classified as recurrent nasopharyngeal carcinoma (rNPC) cases (biopsy-, functional imaging-, or radiological 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. For lesions that were not accessible, the diagnoses were made based on radiologic images. For cases 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) referring to the five-point scale proposed by Ng et al..[34,35] 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 the SYSUCC referring to radiologic criteria for rNPC proposed by Chong et al.[36]

In cases of non-rNPC, patients remained CR without evidence of recurrence within a 5-year follow-up period.

### MR scanning protocol

The follow-up H&N MR examinations were performed on a 1.0-, 1.5-, or 3.0-T MRI unit (SYSUCC: T1-weighted fast spin echo [FSE] images and contrast-enhanced [CE] T1-weighted FSE images on the axial and sagittal planes, CE T1-weighted fat suppressed [FS] images on the coronal plane, and T2-weighted FSE images on the axial plane; CHUCAS & ACHGMU: T1-weighted FSE images on the axial and sagittal planes, CE T1-weighted FS images on the axial and coronal planes, and T2-weighted FS images on the axial plane) (Table S5). All H&N MR images were stored in Digital Imaging and Communications in Medicine (DICOM) format in the imaging database. All MR sequences on different planes are available for radiologists' review.

### MR scan labeling

The H&N MR scans and clinical information (sex, age, primary tumor, first-course of treatment, follow-up data) were retrieved from the Hospital Information System (HIS) in three centers and processed in SYSUCC. All scans with clinical information were evaluated at the individual level by a rotating panel of four board-certified experts (two radiologists and two oncologists) specialized in NPC following the review process depicted in Figure S3. A follow-up H&N MR scan was labeled positive or negative for recurrence (Table S6). The image labels were finalized only when experts reached a consensus.

### MR registration and network architectures

Axial CE T1-weighted imaging (T1-WI) images were collected from the first posttreatment scan identified with radiologic CR until scan with recurrence or the last scan without recurrence within a 5-year follow-up period. Rigid registration between images was performed using voxel-based advanced normalization tools.[37] The current and two previous axial CE T1-WI images were set as target images, and source images, respectively. All images were resampled to a spatial resolution of 0.5 mm × 0.5 mm×6mm.

The input of the model is formed by the three most recent scans, which include the current and two previous scans. If there were fewer than three timepoints, we padded the input with the current scan to reach three. The two-dimension (2D) image slices of the same position in the time-series MR images are superimposed to construct a three-channel image. The input three-channel image consists of 30 slices, covering the structures of the head and neck. To meet the input requirements of the model, all MR images were sampled using bilinear interpolation to a size of 448 × 448 pixels. The input data shape is (3, 30, 448, 448). The imaging features were extracted and fused by an end-to-end architecture that intergrated by the ResNet18[38] (feature extraction module) and Transformer[39] (feature fusion module) (Figure S4). The generated feature maps were divided into feature patches and further projected into the embedding space by the fully connected layer. The Transformer mined the correlations between spatiotemporal feature information and outputted a probability of recurrence likelihood (RAIN score).

We utilized PyTorch with 2 GPUs to train model. The parameter settings were summarized in Table S7. The 'Adam' optimizer optimized the asymmetric loss[40] with minibatch size of 6 and the initial learning rate of $1 \times 10^{-5}$. The probability margin, positive and negative focusing parameters of asymmetric loss were 0.1, 1 and 1.44, respectively. The RandomAffine (rotation, scaler, translation)[41] augmented images during training. The ImageNet pre-trained weights initialized the ResNet18.

### Establishment of triage-driven semiautomated workflow

The RAIN score was a decimal number between 0 and 1, where 1 represented the highest level of suspicion. The score on the scan level was defined by the maximum of the image-level prediction scores.

Two thresholds are set to triage scans into three tiers (Figure 1B): scans scored below 'the rule-out threshold' were considered high-confidence negative (tier 1) in the 'no radiologist workstream'; scans scored above 'the alert threshold' were considered high-confidence positive (tier 2) in the 'enhanced assessment workstream'; scans with scores in between were considered equivocal cases (tier 3) in the 'manual read workstream'.

We used a cumulative substitution scheme to generate two thresholds on the tuning cohort while retaining the accuracy of the original full manual review according to documented radiologic reports (sensitivity: 87.3%; specificity: 97.5%). Scans were sorted from lowest to highest RAIN score. To set 'the rule-out threshold', we determined the maximum deciles of negatives (scans without recurrence) that can be substituted by full automated review by RAIN while retaining undamaged sensitivity to that of original radiologists. To set 'the alert threshold', we determined the maximum percentiles of positives (scans with recurrence) that were preselected by full automated review by RAIN while retaining undamaged specificity to that of original radiologists.

### Reader study

A reader study including 248 eligible cases collected between January 1 and December 31, 2021, was conducted to compare the diagnostic performance of RAINMAN with that of three radiologists (with 30, 19, and 8 years of experience) in SYSUCC. Readers were independent and not involved in labeling the scans. The original radiological interpretations, later recognized labels and case-identified information were removed from scans. The follow-up H&N MR scans were presented in chronological order. Readers reported a judgment of recurrence or nonrecurrence in every single scan (Figure S3). Then, the workflow was simulated in the reader study and compared with the readers in terms of differential diagnostic performance.

## QUANTIFICATION AND STATISTICAL ANALYSIS

Categorical variables are presented as frequencies and percentages. Quantitative statistics are presented as the mean ± SD. The receiver operating characteristic (ROC) curves were created by plotting the proportion of true positives (sensitivity) against the proportion of false positives (1-specificity) by varying the predictive probability threshold. We used the ROC-AUC to show the diagnostic ability in differentiating recurrence scans. A larger ROC-AUC indicated better diagnostic performance. The ROC-AUCs were compared by Delong's test. The threshold value for dichotomy was set by leveraging the Youden method on the ROC curve in the tuning cohort. The diagnostic sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) for identifying targets were evaluated and compared by using the McNemar test, and the 95% CI was calculated using the Clopper-Pearson method. Cohen's kappa agreement analysis was performed on readers and the simulated semiautomated workflow in the reader study.

Statistical analyses were performed using Python (version 3.6.10), R software (version 3.5.1), SPSS statistical software (version 26.0.0.0) and GraphPad Prism software (version 9.0.2). A two-sided $p < 0.05$ was considered significant.

## ADDITIONAL RESOURCES

This semiautomated workflow (RAINMAN) was constructed post hoc based on a retrospective, observational study (ChiCTR.org.cn, ChiCTR2200056595) in three centers from China (Sun Yat-sen University Cancer Center, Guangzhou; Cancer Hospital of The University of Chinese Academy of Sciences, Hangzhou; The Affiliated Cancer Hospital of Guangzhou Medical University, Guangzhou). The Institutional Review Board of Sun Yat-sen University Cancer Center approved this study (B2021-333-01) and performed according to the Helsinki declaration. The requirement for informed consent were waived because of the retrospective and observational design.