Do MRI Radiomic Models Truly Generalize? External Validation Of Three Studies In Parotid Lesion Characterization

Rayan Benyoucef*[1], Martin Goubet*[1], Antoine Barrat*[2], Benoit Chauveau[1], Constance Hordonneau[1], Adrien Bartoli[3,5], Géraud Forestier[1,3], Nicolas Saroul[2,4], Benoît Magnin[1,3,5,Ψ]

*: authors contributed equally

1: Department of Radiology, Clermont University Hospital, France
2: Otolaryngology—Head and Neck Surgery Department, Clermont-Ferrand University Hospital, Clermont-Ferrand, France
3: Institut Pascal, UMR 6602 CNRS, Université Clermont Auvergne, Clermont-Ferrand, France
4: University of Clermont Auvergne, CHU-Clermont-Ferrand, INRAE, UNH, Clermont-Ferrand, France
5: DIA2M, DRCI, Clermont University Hospital, Clermont-Ferrand, France

Ψ : corresponding author

Benoît Magnin
Radiology Department CHU Estaing
1 place Lucie Aubrac 63100 Clermont Ferrand
00 33 4 73 75 02 44
bmagnin@chu-clermontferrand.fr

Abstract

Objectives

External validation of six radiomic models published in three studies: two distinguishing benign from malignant lesions (study 1) and four distinguishing pleomorphic adenomas from Warthin's tumors (study 2 and 3).

Materials & Methods

This monocentric retrospective study included 133 patients who underwent MRI before parotid tumor surgery at our center from 2005 to 2022. For study 1, T1 and T2FS images of 109 benign lesions and 21 malignant ones were included. For study 2, T1 and T2FS images of 58 pleomorphic adenomas and 34 Warthin's tumors were included. For study 3, T2 images of 35 pleomorphic adenomas and 16 Warthin's tumors were included. After segmentation and extraction of the radiomics parameters, the radiomics (Radscore) and combined clinical and radiomics (Nomoscore) models from all 3 studies were applied. Performance was also studied after ComBat harmonization for multiple scanners. Performance was studied on all patients and for study 1 and 2 on a sub-group of 58 patients who had undergone their examination on the same MRI machine.

Results

AUCs were 0.540/0.548 (Radscore/Nomoscore) for study 1, 0.521/0.521 for study 2, and 0.639/0.630 for study 3, whereas the AUCs in the original studies were 0.908/0.938, 0.902/0.918 and 0.796/0.934 respectively. The results were similar after ComBat

harmonization. In the subgroup analysis, the AUCs were 0.533/0.538 for study 1 and 0.513/0.516 for study 2.

Conclusion

Our external validation study was unable to reproduce the results of the six published radiomic models for characterizing parotid lesions, suggesting limited applicability of these radiomic tools in clinical practice.

Key points and Clinical Relevance Statement

Question We aimed to perform an external validation of six previously published MRI radiomic models for the characterization of parotid lesions.

Findings The performances on our population of the six radiomic models were lower than in the initial studies, the highest AUC being 0.639.

Clinical Relevance Statement Our study failed to replicate the performance of the six previously published MRI radiomic models for the characterization of parotid lesions, indicating that the clinical applicability of these radiomic approaches is limited.

Keywords: Radiomic; Parotid gland; Warthin's tumor; Pleomorphic adenoma; External validation

Abbreviations

AUC: area under the curve

BPGT: benign parotid gland tumor

DLI: Deep Lobe Involvement

ICC: Intraclass Correlation Coefficient

IST: Infiltration of Surrounding Tissues

MPGT: malignant parotid gland tumor

PMA: pleomorphic adenoma

ROI: Region of Interest

RQS: Radiomic Quality Score

T1WI: T1-Weighted Imaging

T2FSWI: T2 Fat-Suppressed Weighted Imaging

T2WI: T2-Weighted Imaging

WT: Warthin's tumor

Introduction

Salivary gland tumors are rare lesions, accounting for approximately 3 to 6% of cervical tumors [1]. The parotid gland is the most frequently affected site, representing nearly 80% of these tumors [2]. Among them, about 20% exhibit malignant behavior [3]. Benign tumors encompass several histological types, with pleomorphic adenomas (PMA) and Warthin's tumor (WT) being the most commonly encountered. It is important to emphasize that the prognosis varies depending on the histological type: PMA, in particular, carry an increased risk of recurrence and malignant transformation, which justifies a specific therapeutic approach compared to other benign lesions [4]. Therefore, an accurate characterization of the lesion is essential in order to define an optimal management strategy.

In the initial diagnostic assessment of salivary gland tumors, MRI is considered the reference imaging modality for the radiological evaluation of parotid tumors [5]. It provides detailed morphological characterization that can rival fine needle aspiration in diagnostic performance, while also allowing precise localization of the tumor within the gland and assessment of its extension to adjacent structures [5, 6].

Radiomics is an image analysis method that enables the extraction of quantitative features for diagnostic, prognostic, and therapeutic purposes [7]. Its development in the context of parotid tumors is primarily driven by the goal of enhancing the diagnostic capabilities of imaging; several models have been developed for the characterization of parotid lesions using various MRI sequences [8–11].

There is currently a gap between the large volume of radiomics research publications and its limited application in routine clinical practice, largely due to a lack of reproducibility in reported results [12, 13]. The causes of this limited reproducibility have been investigated [13,

14], and include factors such as test-retest variability, segmentation variability, differences in radiomic feature extraction platforms, and the risk of overfitting. Recommendations have been proposed to guide the development of radiomic models in order to maximize their reproducibility [15–17]. Nevertheless, it is likely that a significant number of published radiomic models lack reproducibility.

We therefore aimed to conduct an external validation of studies focused on the characterization of parotid nodules using standard MRI sequences. After a systematic literature search, three studies were selected. The first study by Zheng et al. used T1-Weighted Imaging (T1WI) and T2 Fat-Suppressed-Weighted Imaging (T2FSWI) combined or not with clinical features to differentiate malignant from benign lesions, achieving a maximum AUC of 0.938 in the validation cohort [9] (referred as study 1). The second study by Zheng et al. also used T1WI and T2FSWI combined or not with clinical features to distinguish pleomorphic adenomas (PMA) from Warthin's tumors (WT), reporting a maximum AUC of 0.918 in the validation cohort [10] (study 2). The third study by Hu et al. used T2-Weighted Imaging (T2WI) combined or not with clinical features to differentiate PMA from WT, with a maximum AUC of 0.934 in the validation cohort [18] (study 3).

The primary objective of this study was to assess the performance of those six MRI-based radiomic models for the parotid lesions characterization on our patients, thereby performing an external validation. The secondary objective were to:

- assess the performance of these models on a homogeneous subgroup of patients whose images were acquired using the same 3T scanner (Discovery, GE Healthcare) to minimize technical noise.

- evaluate the impact of feature harmonization using the ComBat method on model performance, in order to mitigate the bias associated with inter-scanner variability across the entire cohort.

Materials and methods

This study was approved by local Ethics Committee which waived patients' prior written consent due to the retrospective design of the study in compliance to the national policy of individual data protection.

Study selection

All studies investigating the classification of parotid gland lesions using MRI-based radiomics until 31st January 2024 were screened. Studies were then selected guided by three main criteria: the availability of the models' formula, the availability of sufficient local data and the overall radiomic quality. A comprehensive description of the selection is provided in Electronic Supplementary Material 1. A total of 3 studies were finally selected [9, 10, 18].

Methodology

All the steps were reproduced according to the method published in the studies. In case of missing or ambiguous information, the corresponding authors were contacted to obtain details. As no response was obtained after 4 solicitations, the most plausible method was used, and its implementation detailed in the Methods section.

The Radiomic Quality Score [19] et the METRICS [20] of the studies were assessed in consensus by a radiology resident (R.B.) and a senior radiologist with 6 years' experience in radiomics (B.M.).

Patients

We conducted a retrospective, single-center study in which we reviewed all available MRI scans of patients treated for parotid tumors with histological confirmation at our center from October 2005 to February 2022. A total of 204 patients were screened. Among these, 170 patients had benign lesions and 34 had malignant lesions.

The inclusion and exclusion criteria were aligned with those of the respective reference studies. For study 1, inclusion criteria required the presence of complete clinical data and a confirmed diagnosis of either benign or malignant parotid tumor. Exclusion criteria included lesions with a short-axis diameter of less than 5 mm [9]. For studies 2 and 3, the inclusion criteria required complete clinical data and a confirmed diagnosis of pleomorphic adenoma (PMA) or Warthin tumor (WT). The same exclusion criteria as in study 1 applied [10, 18]. Additionally, the availability of the MRI sequences used in the original studies was required: T1 and T2FS sequences for studies 1 and 2, and T2 sequences for study 3. The full flowchart is shown in Figure 1.

MRI acquisitions

The images were acquired using five different MRI scanners from various hospitals within the same university hospital center. However, a subgroup of fifty-seven patients underwent an MRI using the same 3T MRI scanner (Discovery, GE Healthcare). A secondary analysis was conducted on this specific subgroup. The acquisition parameters used in our study and the original studies are detailed in Table 1.

Image segmentation

Blinded to the patients' clinical outcomes, one radiology resident (R.B.) and a senior radiologist with 6 years' experience (B.M.) in parotid imaging performed in consensus the 3D segmentation of the ROI using 3D Slicer software (version 5.6.1; https://www. slicer.org). The tumors were manually delineated at their outermost boundaries slice by slice on each acquisition. The adjacent normal tissue and vessels were not covered. One segmentation only was performed, as the segmentation variability has been already evaluated in the initial articles.

Analysis of radiological characteristics of lesions

The data used to calculate the nomogram in all three studies were collected by a radiology resident (R.B.) who, blinded to the final diagnosis, assessed deep lobe involvement and adjacent tissue infiltration according to the definitions provided in study 1. For study 3, involvement of the parotid tail was also evaluated. In the absence of definition in the article and of answer from the original authors, the definition "the portion that overlies the angle of the mandible" was adopted for the parotid tail.

Image preprocessing, and radiomics feature extraction for studies 1 and 2

Image preprocessing was performed as described by authors: "N4ITK" bias field correction was applied [21] and the "$\mu \pm 3\ \sigma$" method was used to correct for the effects of different MR scanners and acquisition protocols and normalize the image intensities [22]. Radiomic feature extraction was performed using PyRadiomics [23] (version 3.0.1) with a 1 × 1 × 1 mm3 resampling to extract the same 1702 3D radiomic features on T1WI and T2FSWI. The details of the discretization were not mentioned in the article. As no answer was obtained from the

authors, a fixed bin width of 25 and a sitkBSpline interpolator were used, as they are the default parameter in the 3DSlicer PyRadiomics interface. A Z-score normalization was then performed.

Image preprocessing, and radiomics feature extraction for study 3

A resampling with an arbitrary size of 1 × 1 × 1 mm3 was applied, as the voxel size for resampling was not mentioned in the article and no answer was obtained from the authors. No other image preprocessing was performed. Radiomic feature extraction was performed using PyRadiomics [23] (version 3.0.1) to extract the same 971 3D radiomic features on T2WI, and Z-score normalization was then performed. The same extraction parameters as in studies 1 and 2 were used.

Models

For each study, a Radscore (score derived from radiomic features to evaluate the diagnosis) and a Nomoscore (score derived from Radscore and clinical variables) were built as in the initial studies.

Models for studies 1 and 2

Radscore and Nomoscore were built using the same combination of selected radiomic features and/or clinical features (Electronic Supplementary Material 2 and 3) as described by the authors.

Models for study 3

Radscore was built using the same combination of selected radiomic features (Electronic Supplementary Material 4) as described by the authors. As no formula was mentioned for the radiomics-clinical model and no answer was obtained from the authors, the formula for the

radiomics-clinical model was estimated after measurements on the graphical nomogram. The equations used for this study are:

Points = 141.6811 + 15.29 * radscore + 1.111111 * age + 17.7 * (1-parotid tail)

Risk (Nomoscore) = 0.02781 * Points - 2.01682

A table describing the methods described in the original articles, the missing information and the methods applied in our study is shown in Table 2. A table with the main results of the 6 models in the original articles is shown in Table 3.

ComBat Harmonization

To mitigate the potential bias arising from the use of different MRI scanners in our study, we replicated the results applying the ComBat harmonization method to the radiomic features [24].

Statistical analyzes

A Shapiro-Wilk normality test was performed on the data to justify the use of non-parametric tests. Differences in quantitative variables were assessed using the Wilcoxon test, while differences in categorical variables were evaluated using the χ^2 test.

Subsequently, area under the curve (AUC), sensitivity, specificity, and accuracy were calculated to assess the diagnostic performance of the models. Calibration curves and Brier scores were also evaluated.

Results

The demographic data, Radscore, and Nomoscore of the patients included for each study are presented in Table 4. The detail of the histologic subtypes for study 1 is presented in Electronic Supplementary Material 5.

The performances of the models in our cohort are presented in Table 5.

The ROC curves along with the corresponding AUCs for models are shown in Figure 2.

The Brier scores and the calibration curves of the models are shown in Electronic Supplementary Material 6. The Brier score are between 0 and 0.2318.

Results after ComBat harmonization

The results of AUC for all studies after ComBat harmonization are shown in Figure 3 and in Table 5.

Subgroup analysis using the same MRI scanner

Study 1

57 patients who underwent MRI using the same scanner model (3T Discovery, GE Healthcare) with T1WI and T2FSWI were included in this subgroup analysis for study 1. Among these patients, 49 had benign lesions and 8 had malignant lesions. The performance of the models within this subgroup is presented in Figure 4A and Table 5.

Study 2

Among the 57 patients who underwent imaging on the same 3T Discovery MRI scanner from GE Healthcare, 27 were diagnosed with PMA and 13 with WT. These patients were included

in the subgroup analysis for study 2. The performance of the models in this subgroup is presented in Figure 4B and Table 5.

Study 3

There were not enough patients to perform a subgroup analysis.

Radiomic Quality Score and METRICS

The Radiomic Quality Score (RQS) was 13 out of 36 (36.11%) for study 1, 13 out of 36 (36.11%) for study 2, and 12 out of 36 (33.33y%) for study 3. The METRICS was 74.9% for studies 1 and 2 and 58.6% for study 3. (see Electronic Supplementary Material 7)

Discussion

Our objective was to perform an external validation of six radiomic models: two designed to distinguish malignant from benign parotid tumors, and four others to distinguish PMA from WT. The AUCs of the models combining clinical and radiomic features in our population were 0.548, 0.521, and 0.630, compared to 0.938, 0.918, and 0.934 in the original studies. Similar results were observed when evaluating radiomic models alone.

Thus, our study did not validate any of the published radiomic models. In detail, none of the scores significantly differed between the groups to be distinguished, unlike in studies 1 and 2. Some scores even showed markedly different magnitudes - for instance, mean Nomoscores of study 1 were 31.51 and 35.54 in our study compared to −2.305 and 7.830 in the validation cohort of the original article. However, certain clinical parameters mirrored those reported in the original studies: malignant lesions were associated with older age, deeper lobe involvement, and adjacent tissue infiltration compared to benign ones; WT were more frequent in male patients compared to PMA. Nevertheless, we did not observe the differences

in age or parotid tail involvement when distinguishing between PMA and WT reported in studies 2 and 3.

The weakness of the results leads us to fear methodological anomalies in the creation of these radiomics models. Yet, certain aspects of the original methodology did adhere to recommendations aimed at improving reproducibility [13, 14]. A second segmentation was performed to retain only radiomic features with satisfactory inter- and intra-observer reproducibility. An ICC threshold of 0.75 was applied to select reproducible features, which is commonly used; however, adopting a more stringent threshold of 0.9 could further ensure greater reproducibility [25]. Data reduction and model construction were carried out using conventional approaches. The developed models were generally well-documented, except for the mixed clinical and radiomic model in study 3, which was only presented graphically, which represents is a clear limitation. For study 3, validation was conducted on a subset from the same center. In contrast, for studies 1 and 2, external validation was performed using data from a center different from the training images. This type of external validation is typically considered essential to confirm that the model is not overfitted to the original dataset.

Conversely, several methodological factors likely contributed to this poor reproducibility. Manual segmentation, as used here, is inherently less reproducible than semi-automated methods [26]. Furthermore, the lack of normalization in study 3 [27] and missing technical details - such as undisclosed discretization or resampling parameters - hinder replication. These omissions underscore the high sensitivity of radiomic features to preprocessing variations and the critical impact of incomplete reporting. Similarly, the definition of the parotid tail was not specified. In the absence of clarification from the authors, the most commonly used definitions and methods were applied in this study. These points highlight the

importance of a detailed technical reporting in radiomics publications. These undocumented deviations from the original pipelines likely contribute to the performance drop, further complicating the path toward clinical implementation. Moreover, the use of the PyRadiomics platform may introduce reproducibility issues, even though it is among the most widely used platforms [28]. Lastly, the models relied predominantly on texture features, which are generally less reproducible than first-order features [27, 29].

Our results suggest significant overfitting in the original models, which failed to replicate their initial performance. Despite reported AUCs exceeding 0.9, we observed a drop to 0.521–0.639, indicating a total loss of discriminative power. Brier scores (up to 0.232) confirm poor calibration for our population; though below the 0.25 'non-informative' threshold, they reflect substantial predictive inaccuracy. This discrepancy suggests that the original radiomic features are likely center-specific or overfitted, severely limiting their generalizability to different clinical settings.

A scanner or protocol variability might contribute to the model failure. Indeed, we evaluated them on a cohort imaged using five different MRI scanners, which reflects routine clinical diversity. Nevertheless, analysis of the first two models in a subgroup examined on the same MRI scanner yielded similarly poor results, and the results were similar when performing ComBat harmonization for scanner heterogeneity, suggesting a limited influence of scanner variability in the failure.

Beyond technical factors, a significant population shift likely influenced model performance. Notably, the malignancy prevalence in study 1 (48%) far exceeds our cohort (16%), which aligns better with European data [3]. While the impact of ethnicity on MRI appearance remains undocumented, differences in recruitment criteria and disease distribution (16% vs.

48%) create a major bias. Consequently, these models lack generalizability to our clinical setting. Furthermore, the small size of certain subgroups - such as the 16 cases for study 3 - requires cautious interpretation of our AUCs (0.521–0.639). These findings underscore the difficulty of maintaining stable performance when transitioning from curated, balanced development sets to heterogeneous, real-world clinical cohorts.

Although the results are disappointing, they do not call into question the rigor of the original studies. Indeed, the RQS scores of the studies range from 33.3% to 36.1%, which, while modest, are above the average reported for studies from slightly earlier periods, with a mean RQS of 26.1% [30]. Moreover, the METRICS of studies 1 and 2 are evaluated as good, the METRICS of study 3 being moderate. This highlights concerns about the external validity of other studies with even lower RQS or METRICS.

Several studies have aimed to externally validate published radiomic models [31–33], but only a few have reported negative findings [34–36]. We believe that this type of study underscores the importance of adhering to a very high methodological standard [15–17] in the development of radiomic models. Our study also highlights the urgent need for adherence to standardized reporting guidelines and mandatory code sharing to ensure the transparency of radiomic models. Without such rigor, external validation - and therefore clinical applicability - may not be achievable.

Our external validation study of six radiomic models - two distinguishing malignant from benign parotid tumors and four differentiating pleomorphic adenomas from Warthin tumors - did not succeed in reproducing the original results. This work highlights the necessity of strict methodological rigor and external validation of radiomic models prior to their clinical application.

References

1. Guzzo M, Locati LD, Prott FJ, et al (2010) Major and minor salivary gland tumors. Crit Rev Oncol Hematol 74:134–148. https://doi.org/10.1016/j.critrevonc.2009.10.004

2. Gao M, Hao Y, Huang MX, et al (2017) Salivary gland tumours in a northern Chinese population: a 50-year retrospective study of 7190 cases. Int J Oral Maxillofac Surg 46:343–349. https://doi.org/10.1016/j.ijom.2016.09.021

3. Bradley PJ, McGurk M (2013) Incidence of salivary gland neoplasms in a defined UK population. British Journal of Oral and Maxillofacial Surgery 51:399–403. https://doi.org/10.1016/j.bjoms.2012.10.002

4. Valstar MH, de Ridder M, van den Broek EC, et al (2017) Salivary gland pleomorphic adenoma in the Netherlands: A nationwide observational study of primary tumor incidence, malignant transformation, recurrence, and risk factors for recurrence. Oral Oncol 66:93–99. https://doi.org/10.1016/j.oraloncology.2017.01.004

5. Coudert H, Mirafzal S, Dissard A, et al (2021) Multiparametric magnetic resonance imaging of parotid tumors: A systematic review. Diagnostic and Interventional Imaging 102:121–130. https://doi.org/10.1016/j.diii.2020.08.002

6. Vergez S, Fakhry N, Cartier C, et al (2021) Guidelines of the French Society of Otorhinolaryngology-Head and Neck Surgery (SFORL), part I: Primary treatment of pleomorphic adenoma. Eur Ann Otorhinolaryngol Head Neck Dis 138:269–274. https://doi.org/10.1016/j.anorl.2020.09.002

7. Gillies RJ, Kinahan PE, Hricak H (2016) Radiomics: Images Are More than Pictures, They Are Data. Radiology 278:563–577. https://doi.org/10.1148/radiol.2015151169

8. Faggioni L, Gabelloni M, De Vietro F, et al (2022) Usefulness of MRI-based radiomic features for distinguishing Warthin tumor from pleomorphic adenoma: performance assessment using T2-weighted and post-contrast T1-weighted MR images. Eur J Radiol Open 9:100429. https://doi.org/10.1016/j.ejro.2022.100429

9. Zheng Y, Li J, Liu S, et al (2020) MRI-Based radiomics nomogram for differentiation of benign and malignant lesions of the parotid gland. Eur Radiol. https://doi.org/10.1007/s00330-020-07483-4

10. Zheng Y, Chen J, Xu Q, et al (2021) Development and validation of an MRI-based radiomics nomogram for distinguishing Warthin's tumour from pleomorphic adenomas of the parotid gland. Dentomaxillofac Radiol 50:20210023. https://doi.org/10.1259/dmfr.20210023

11. Gabelloni M, Faggioni L, Attanasio S, et al (2020) Can Magnetic Resonance Radiomics Analysis Discriminate Parotid Gland Tumors? A Pilot Study. Diagnostics 10:900. https://doi.org/10.3390/diagnostics10110900

12. Traverso A, Wee L, Dekker A, Gillies R (2018) Repeatability and Reproducibility of Radiomic Features: A Systematic Review. International Journal of Radiation Oncology*Biology*Physics 102:1143–1158. https://doi.org/10.1016/j.ijrobp.2018.05.053

13. Pinto Dos Santos D, Dietzel M, Baessler B (2021) A decade of radiomics research: are images really data or just patterns in the noise? Eur Radiol 31:1–4. https://doi.org/10.1007/s00330-020-07108-w

14. Zwanenburg A (2019) Radiomics in nuclear medicine: robustness, reproducibility, standardization, and how to avoid data analysis traps and replication crisis. Eur J Nucl Med Mol Imaging 46:2638–2655. https://doi.org/10.1007/s00259-019-04391-8

15. Zwanenburg A, Vallières M, Abdalah MA, et al (2020) The Image Biomarker Standardization Initiative: Standardized Quantitative Radiomics for High-Throughput Image-based Phenotyping. Radiology 295:328–338. https://doi.org/10.1148/radiol.2020191145

16. Zwanenburg A, Leger S, Vallières M, Löck S (2019) Image biomarker standardisation initiative

17. Santinha J, Pinto Dos Santos D, Laqua F, et al (2024) ESR Essentials: radiomics—practice recommendations by the European Society of Medical Imaging Informatics. Eur Radiol. https://doi.org/10.1007/s00330-024-11093-9

18. Hu Z, Guo J, Feng J, et al (2022) Value of T2-weighted-based radiomics model in distinguishing Warthin tumor from pleomorphic adenoma of the parotid. Eur Radiol 33:4453–4463. https://doi.org/10.1007/s00330-022-09295-0

19. Lambin P, Leijenaar RTH, Deist TM, et al (2017) Radiomics: the bridge between medical imaging and personalized medicine. Nature Reviews Clinical Oncology 14:749–762. https://doi.org/10.1038/nrclinonc.2017.141

20. Kocak B, Akinci D'Antonoli T, Mercaldo N, et al (2024) METhodological RadiomICs Score (METRICS): a quality scoring tool for radiomics research endorsed by EuSoMII. Insights Imaging 15:8. https://doi.org/10.1186/s13244-023-01572-w

21. Tustison NJ, Avants BB, Cook PA, et al (2010) N4ITK: Improved N3 Bias Correction. IEEE Trans Med Imaging 29:1310–1320. https://doi.org/10.1109/TMI.2010.2046908

22. Collewet G, Strzelecki M, Mariette F (2004) Influence of MRI acquisition protocols and image intensity normalization methods on texture classification. Magnetic Resonance Imaging 22:81–91. https://doi.org/10.1016/j.mri.2003.09.001

23. van Griethuysen JJM, Fedorov A, Parmar C, et al (2017) Computational Radiomics System to Decode the Radiographic Phenotype. Cancer Res 77:e104–e107. https://doi.org/10.1158/0008-5472.CAN-17-0339

24. Orlhac F, Lecler A, Savatovski J, et al (2021) How can we combat multicenter variability in MR radiomics? Validation of a correction procedure. Eur Radiol 31:2272–2280. https://doi.org/10.1007/s00330-020-07284-9

25. Jha AK, Mithun S, Jaiswar V, et al (2021) Repeatability and reproducibility study of radiomic features on a phantom and human cohort. Sci Rep 11:2055. https://doi.org/10.1038/s41598-021-81526-8

26. Rios Velazquez E, Aerts HJWL, Gu Y, et al (2012) A semiautomatic CT-based ensemble segmentation of lung tumors: comparison with oncologists' delineations and with the surgical specimen. Radiother Oncol 105:167–173. https://doi.org/10.1016/j.radonc.2012.09.023

27. Veiga-Canuto D, Fernández-Patón M, Cerdà Alberich L, et al (2024) Reproducibility Analysis of Radiomic Features on T2-weighted MR Images after Processing and Segmentation Alterations in Neuroblastoma Tumors. Radiology: Artificial Intelligence 6:e230208. https://doi.org/10.1148/ryai.230208

28. Bettinelli A, Marturano F, Avanzo M, et al (2022) A Novel Benchmarking Approach to Assess the Agreement among Radiomic Tools. Radiology 303:533–541. https://doi.org/10.1148/radiol.211604

29. Rai R, Holloway LC, Brink C, et al (2020) Multicenter evaluation of MRI-based radiomic features: A phantom study. Medical Physics 47:3054–3063. https://doi.org/10.1002/mp.14173

30. Park JE, Kim D, Kim HS, et al (2020) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 30:523–536. https://doi.org/10.1007/s00330-019-06360-z

31. Mori M, Deantoni C, Olivieri M, et al (2023) External validation of an 18F-FDG-PET radiomic model predicting survival after radiotherapy for oropharyngeal cancer. Eur J Nucl Med Mol Imaging 50:1329–1336. https://doi.org/10.1007/s00259-022-06098-9

32. Bos P, Martens RM, de Graaf P, et al (2023) External validation of an MR-based radiomic model predictive of locoregional control in oropharyngeal cancer. Eur Radiol 33:2850–2860. https://doi.org/10.1007/s00330-022-09255-8

33. Cusumano D, Boldrini L, Yadav P, et al (2021) Delta radiomics for rectal cancer response prediction using low field magnetic resonance guided radiotherapy: an external validation. Phys Med 84:186–191. https://doi.org/10.1016/j.ejmp.2021.03.038

34. Foley KG, Shi Z, Whybra P, et al (2019) External validation of a prognostic model incorporating quantitative PET image features in oesophageal cancer. Radiotherapy and Oncology 133:205–212. https://doi.org/10.1016/j.radonc.2018.10.033

35. Shahzadi I, Zwanenburg A, Lattermann A, et al (2022) Analysis of MRI and CT-based radiomics features for personalized treatment in locally advanced rectal cancer and external validation of published radiomics models. Sci Rep 12:10192. https://doi.org/10.1038/s41598-022-13967-8

36. van der Reijd DJ, Guerendel C, Staal FCR, et al (2024) Independent validation of CT radiomics models in colorectal liver metastases: predicting local tumour progression after ablation. Eur Radiol 34:3635–3643. https://doi.org/10.1007/s00330-023-10417-5

Tables

Table 1: MRI acquisition parameters in the initial studies and in the current study.

	studies 1 and 2				study 3		
	training set		validation set				
Number of patients	80 (study 1), 75 (study 2)		35 (study 1),52 (study 2)		?	?	?
Manufacturer	GE Healthcare		Siemens		Phillips	Phillips	Siemens
Model	Signa 3T		Skyra 3T		Ingenia 3T	Achieva 1.5T	Skyra 3T
sequence	T1	T2FS	T1	T2FS	T2	T2	T2
Repetition time (ms)	420	3600	500	3000	2500	2818	3000
Echo time (ms)	11	102	11	103	80	89	110
voxel size (mm^3)	0.69 x 0.86 x 4	0.69 x 0.86 x 4	0.69 x 1.08 x 4	0.69 x 1.08 x 4	1.1 x 1.3 x 5	0.6 x 0.71 x 4 or 5	0.63 x 0.63 x 4

	Our study								
Number of patients	57			33			22		
Manufacturer	GE Healthcare			GE Healthcare			Siemens		
Model	Discovery 3T			Optima 1.5T			Aera 1.5T		
sequence	T1	T2FS	T2	T1	T2FS	T2	T1	T2FS	T2
Repetition time (ms)	600	4874	4077	540	6500	5800	520	5300	4890
Echo time (ms)	11	90	92	9	77	101	12	100	100
voxel size (mm^3)	0.47 x 0.47 x 3	0.47 x 0.47 x 3	0.47 x 0.47 x 3	0.43 x 0.43 x 3	0.43 x 0.43 x 3	0.43 x 0.43 x 3	0.52 x 0.52 x 3.5	0.66 x 0.66 x 3.5	0.66 x 0.66 x 3.5

	Our study					
Number of patients	15			6		
Manufacturer	Siemens			GE Healthcare		
Model	Avanto 1.5T			Signa 1.5T		
sequence	T1	T2FS	T2	T1	T2FS	T2
Repetition time (ms)	606	5420	7590	500	3500	3000
Echo time (ms)	12	118	97	10	86	120
voxel size (mm^3)	0.67 x 0.67 x 4	0.6 x 0.6 x 4	0.6 x 0.6 x 4	0.47 x 0.47 x 3	0.47 x 0.47 x 3	0.47 x 0.47 x 3.5

Table 2: Methods described in the original studies, the missing information and the methods applied in our study

study	model	phase	step	information	strategy	performed
study 1 and study 2	nomoscore and radsore	pre processing	bias field correction	complete	replicated	N4ITK bias filed correction
			correction for different MR scanners	complete	replicated	"μ ±3 σ" method
		feature extraction	resampling	complete	replicated	1 × 1 × 1 mm3 resampling
			discretization	not mentioned	Use default parameter in PyRadiomics	fixed bin width of 25 and a sitkBSpline interpolator
			feature normalization	complete	replicated	Z score normalization
		classification	selected features	complete	replicated	see Electronic Supplementary material 2 and 3
			features coefficient	complete	replicated	see Electronic Supplementary material 2 and 3
	nomoscore	clinical variables		complete	replicated	see Electronic Supplementary material 2 and 3
study 3	nomoscore and radsore	pre processing		not mentioned		no image preprocessing was performed
		feature extraction	resampling	Partial: size not mentioned	Use default parameter in PyRadiomics	1 × 1 × 1 mm3 resampling
			discretization	not mentioned	Use default parameter in PyRadiomics	fixed bin width of 25 and a sitkBSpline interpolator
			feature normalization	complete	replicated	Z score normalization
		classification	selected features	complete	replicated	see Electronic Supplementary material 4
			features coefficient	Uncomplete: graphical representation only		coefficient extracted from graphical nomogram
	nomoscore	clinical variables		Uncomplete: unprecise definition of the parotid tail	Use of common definition	"the portion that overlies the angle of the mandible" was adopted for the parotid tail.

Table 3: Summary of study characteristics and main results of the 6 models in the original studies. The features used in the models are detailed in Electronic Supplementary Material 2, 3 and 4.

Study and model	Training Population (n)	Test Population (n)	Type of validation	Training Performance Metrics	Test Performance Metrics
Study 1	**n=80** Benign: 42 Malignant: 38	**n=35** Benign: 18 Malignant: 17	Truly independent external test set		
Radscore (radiomic features only)				AUC: 0.944 Sen: 0.921 Spec:0. 905 Acc: 0.913	AUC: 0.908 Sen: 0.882 Spec :0. 778 Acc: 0.829
Nomoscore (radiomic and clinical features)				**AUC: 0.952** Sen: 0.921 Spec: 0.905 Acc: 0.913	**AUC: 0.938** Sen: 0.941 Spec: 0.833 Acc: 0.886
Study 2	**n=75** WT: 34 PMA: 41	**n=52** WT: 24 PMA: 28	Truly independent external test set		
Radscore (radiomic features only)				AUC: 0.926 Sen: 0.902 Spec: 0.882 Acc: 0.893	AUC: 0.902 Sen: 0.786 Spec: 0.875 Acc: 0.827
Nomoscore (radiomic and clinical features)				**AUC: 0.953** Sen: 0.927 Spec: 0.853 Acc: 0.893	**AUC: 0.918** Sen: 0.893 Spec: 0.833 Acc: 0.865
Study 3	**n=82** WT: 42 PMA: 40	**n=35** WT: 19 PMA: 16	Truly independent test set from same center		
Radscore (radiomic features only)				AUC: 0.826 Sen: 0.900 Spec: 0.691 Acc: 0.793	AUC: 0.796 Sen: 0.563 Spec: 0.947 Acc: 0.771
Nomoscore (radiomic and clinical features)				**AUC: 0.962** Sen: 0.875 Spec: 0.952 Acc: 0.915	**AUC: 0.934** Sen: 0.938 Spec: 0.842 Acc: 0.886

Table 4: Demographic data, Radscore and Nomoscore for original studies. Data are presented as n (%) or mean +/- SD.

		Study 1			Study 2			Study 3		
		benign	malignant	p-value	PMA	WT	p-value	PMA	WT	p-value
n		109	21		59	34		35	16	
gender M	n (%)	54 (49.54%)	10 (47.62%)	**1.00**	16 (27.11%)	29 (85.29%)	**<0.001**	7 (20.00%)	12 (75.00%)	**<0.001**
Age, year	(mean +/- SD)	53.6 +/- 14.3	62.2 +/- 17.2	**0.002**	52.4 +/- 15.6	55.5 +/- 8.91	**0.35**	52.5 +/- 16.3	55.4 +/- 8.16	**0.62**
maximum diameter, mm	(mean +/- SD)	21.8 +/- 9.9	29.7 +/- 13.1	**0.008**	22.4 +/- 10.7	20.4 +/- 8.15	**0.57**	19.6 +/- 8.46	19.6 +/- 8.58	**0.98**
DLI	(absent/present)	79/30	7/14	**0.001**						
IST	(absent/present)	100/9	15/6	**0.02**						
parotid tail	(absent/present)							21/14	6/10	**0.24**
Radscore	Mean	8.23	7.18	**0.49**	-1.41	-1.28	**0.74**	-1.07	-2.96	**0.14**
	SD	5.27	4.85		2.62	2.87		4.77	5.98	
Nomoscore	Mean	35.54	31.51	**0.57**	3.24	3.99	**0.73**	77.58	41.49	**0.11**
	SD	22.47	20.90		10.08	11.35		79.18	88.43	

Table 5: Summary of performances for all models in the original studies and in our study.
Data are presented as value or value [95%CCI]

			Original studies			
			Training population	**Test population**	**No harmonization**	**Co**
Study 1	Radscore	AUC	**0.944 [0.868-0.983]**	**0.908 [0.762-0.979]**	**0.54 [0.411-0.668]**	**0**
		Accuracy	0.913	0.829	0.562 [0.246-0.862]	0
		Sensitivity	0.921	0.882	0.667 [0.143-1]	
		Specificity	0.905	0.778	0.541 [0.101-1]	
	Nomoscore	AUC	**0.962 [0.880-0.987]**	**0.938 [0.801-0.991]**	**0.548 [0.419-0.677]**	**0**
		Accuracy	0.913	0.886	0.569 [0.254-0.854]	0
		Sensitivity	0.921	0.941	0.667 [0.189-1]	
		Specificity	0.905	0.833	0.55 [0.11-0.991]	
Study 2	Radscore	AUC	**0.926 [0.842-0.974]**	**0.902 [0.787-0.967]**	**0.521 [0.395-0.648]**	**(**
		Accuracy	0.893	0.827	0.591 [0.451-0.699]	(
		Sensitivity	0.902	0.786	0.5 [0.059-0.941]	0
		Specificity	0.882	0.875	0.644 [0.153-1]	
	Nomoscore	AUC	**0.953 [0.878-0.989]**	**0.918 [0.808-0.976]**	**0.521 [0.395-0.647]**	**(**
		Accuracy	0.893	0.865	0.581 [0.43-0.72]	0
		Sensitivity	0.927	0.893	0.559 [0.088-0.941]	0
		Specificity	0.853	0.833	0.593 [0.153-0.983]	(
Study 3	Radscore	AUC	**0.826 [0.736-0.916]**	**0.796 [0.646-0.946]**	**0.639 [0.469-0.809]**	**0**
		Accuracy	0.793	0.771	0.647 [0.51-0.804]	
		Sensitivity	0.9	0.563	0.688 [0.25-1]	
		Specificity	0.691	0.947	0.629 [0.343-1]	
	Nomoscore	AUC	**0.962 [0.922-1]**	**0.934 [0.858-1]**	**0.63 [0.454-0.807]**	**0**
		Accuracy	0.915	0.886	0.549 [0.51-0.824]	
		Sensitivity	0.875	0.938	0.875 [0.188-1]	
		Specificity	0.952	0.842	0.4 [0.314-1]	

Figures

Figure 1 Study workflow (PMA: pleomorphic adenoma, WT: Warthin's tumor)

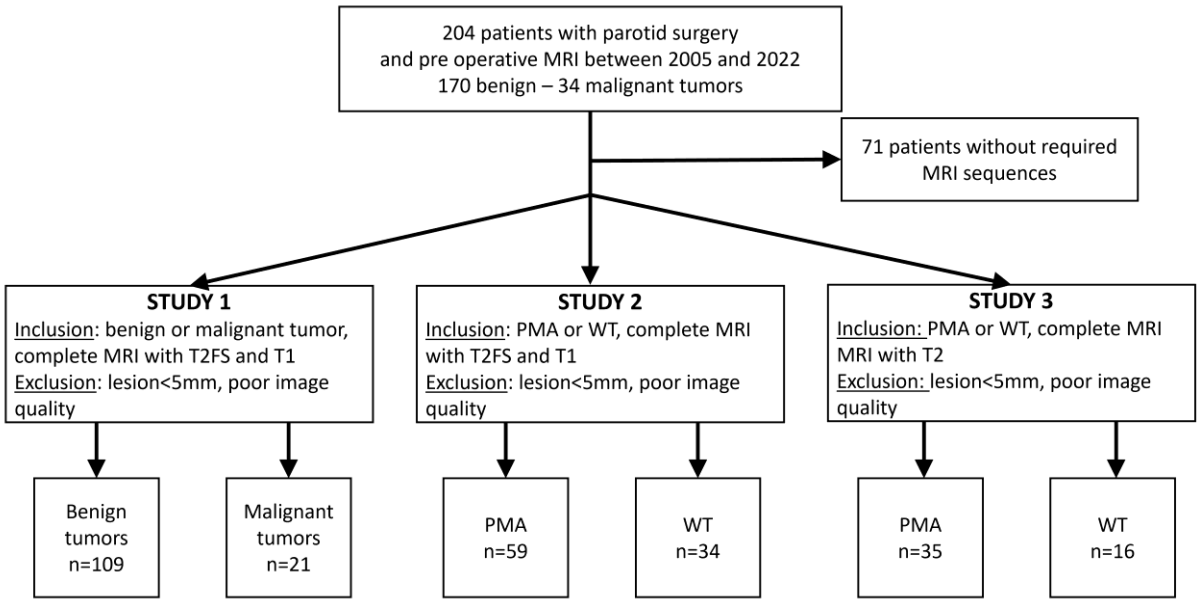


Figure 2 ROC curve and corresponding AUC for Nomoscore and Radscore for all studies.

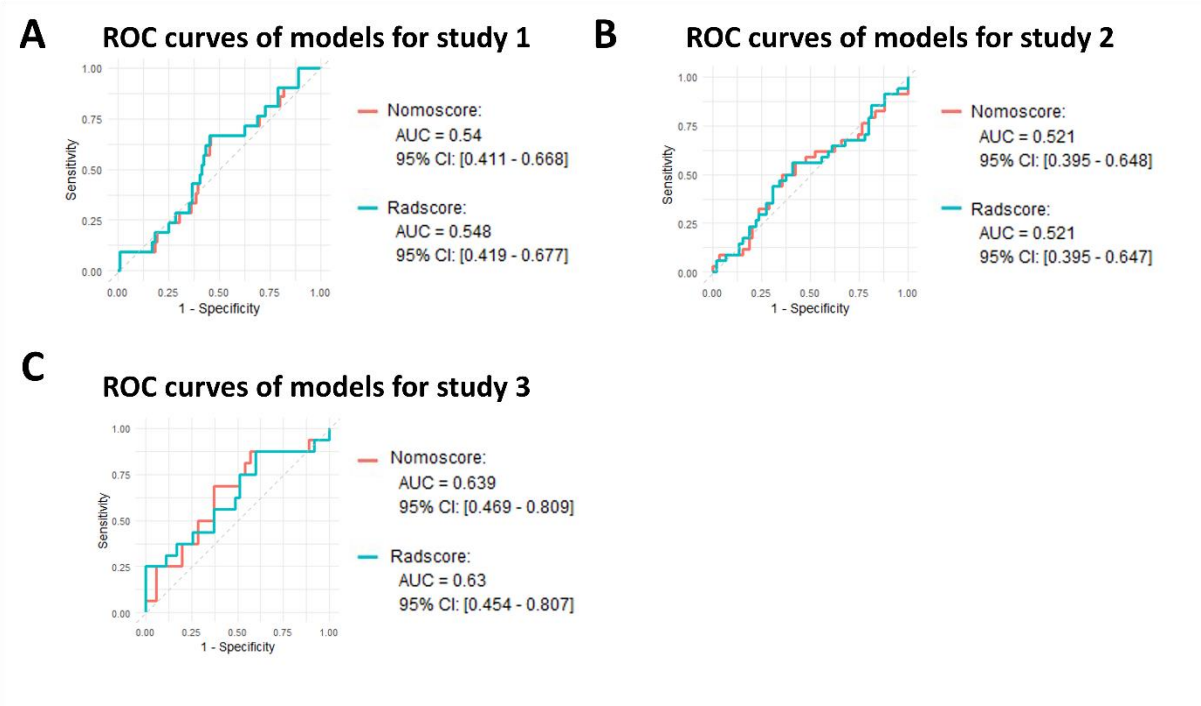


Figure 3: ROC curve and corresponding AUC for Nomoscore and Radscore for all studies after ComBat harmonization

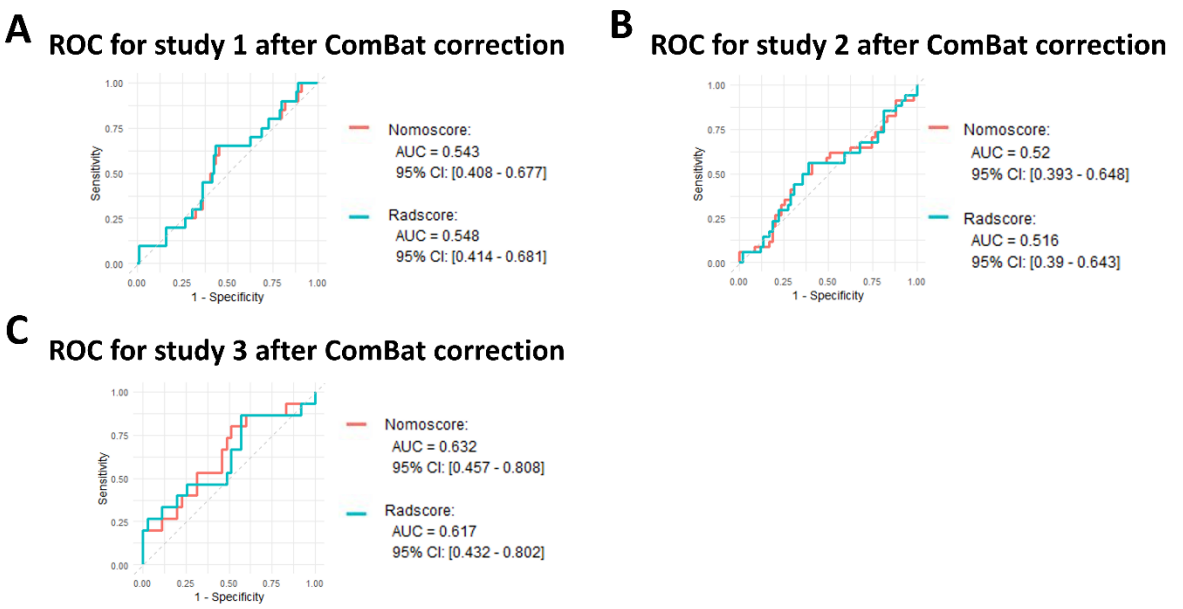
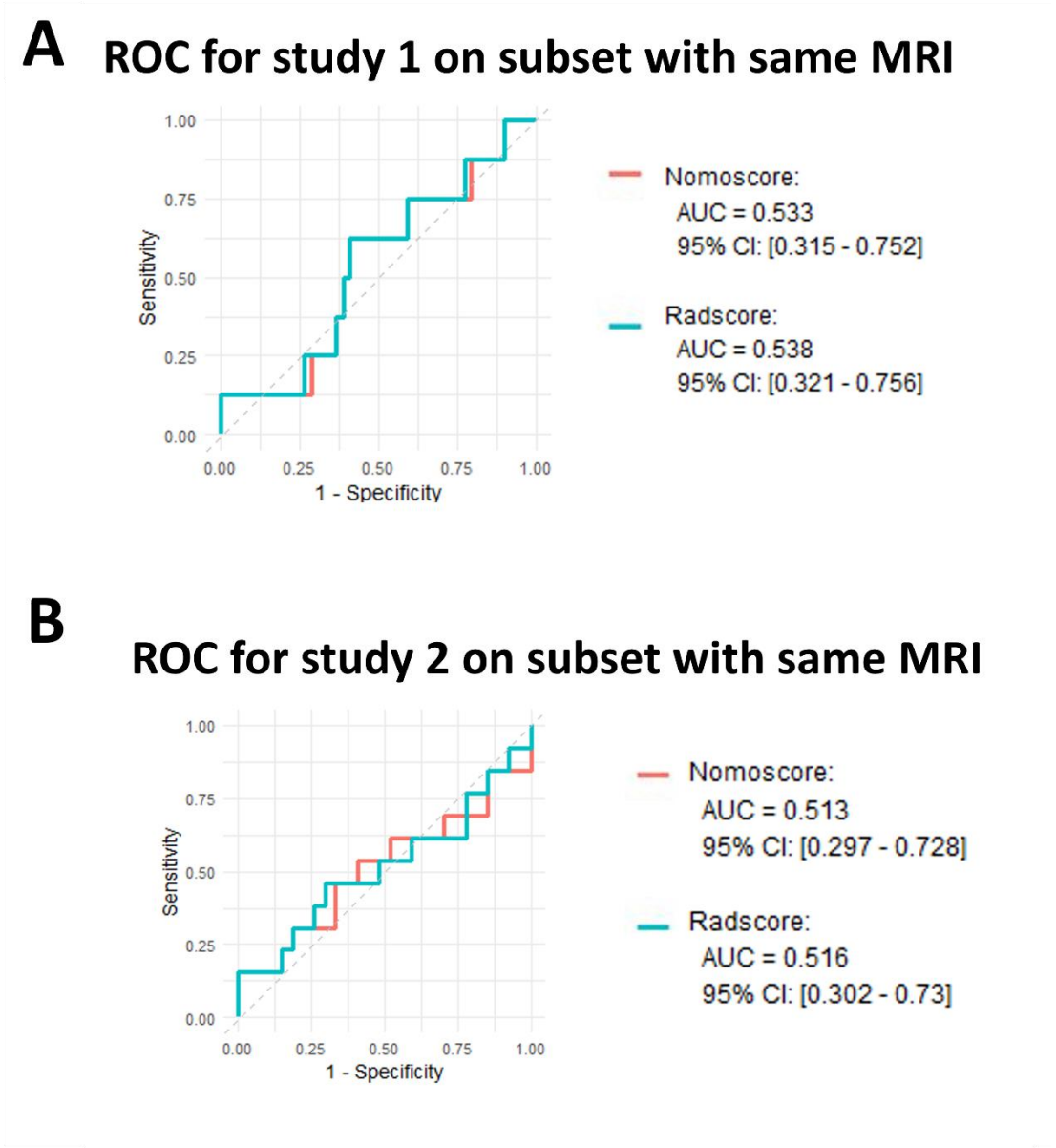


Figure 4: ROC curve and corresponding AUC for Nomoscore and Radscore for all studies on subset of patients on the same MRI



Electronic Supplementary Material

1 Electronic Supplementary Material 1

Article selection process

Systematic PubMed search

A systematic search was conducted on the PubMed database to identify studies investigating the classification of parotid gland lesions using MRI-based radiomics until 31st January 2024. The search string is detailed in the flowchart (Supp Figure 1).

Inclusion and Exclusion Criteria

The selection process was divided into two phases: an initial screening of the literature and a focused technical analysis of recent publications.

Initial Screening

Studies were excluded based on the following criteria:

- study type: review articles, meta-analyses, case reports, and conference abstracts.
- methodology: studies not utilizing radiomics (e.g., qualitative radiological assessment or Deep Learning only).
- clinical objective: studies that did not aim specifically to differentiate:
    - malignant tumours (MT) vs. benign tumours (BT)
    - or pleomorphic adenoma (PMA) vs. Warthin's Tumor (WT)

Technical Refinement

To ensure reproducibility and standardization in our analysis, we further screened the remaining recent articles. Studies were excluded if they met any of the following technical criteria:

- feature extraction software: articles not using recognized, standardized IBSI compliant extraction platforms.
- mathematical transparency: articles where the specific formulas were not available or clearly referenced.
- MRI sequences: studies utilizing sequences other than T1-weighted (T1w), T2-weighted (T2w), or Fat-Suppressed T2-weighted (T2FS) imaging. Indeed, the availability of other sequences in our cohort was not sufficient to ensure sufficient data with other MRI sequences.
- studies with a RQS under 26.1% (which was the mean RQS on published data [1])

The details of the 20 analyzed articles are available in Supp Table 1.

Final selection

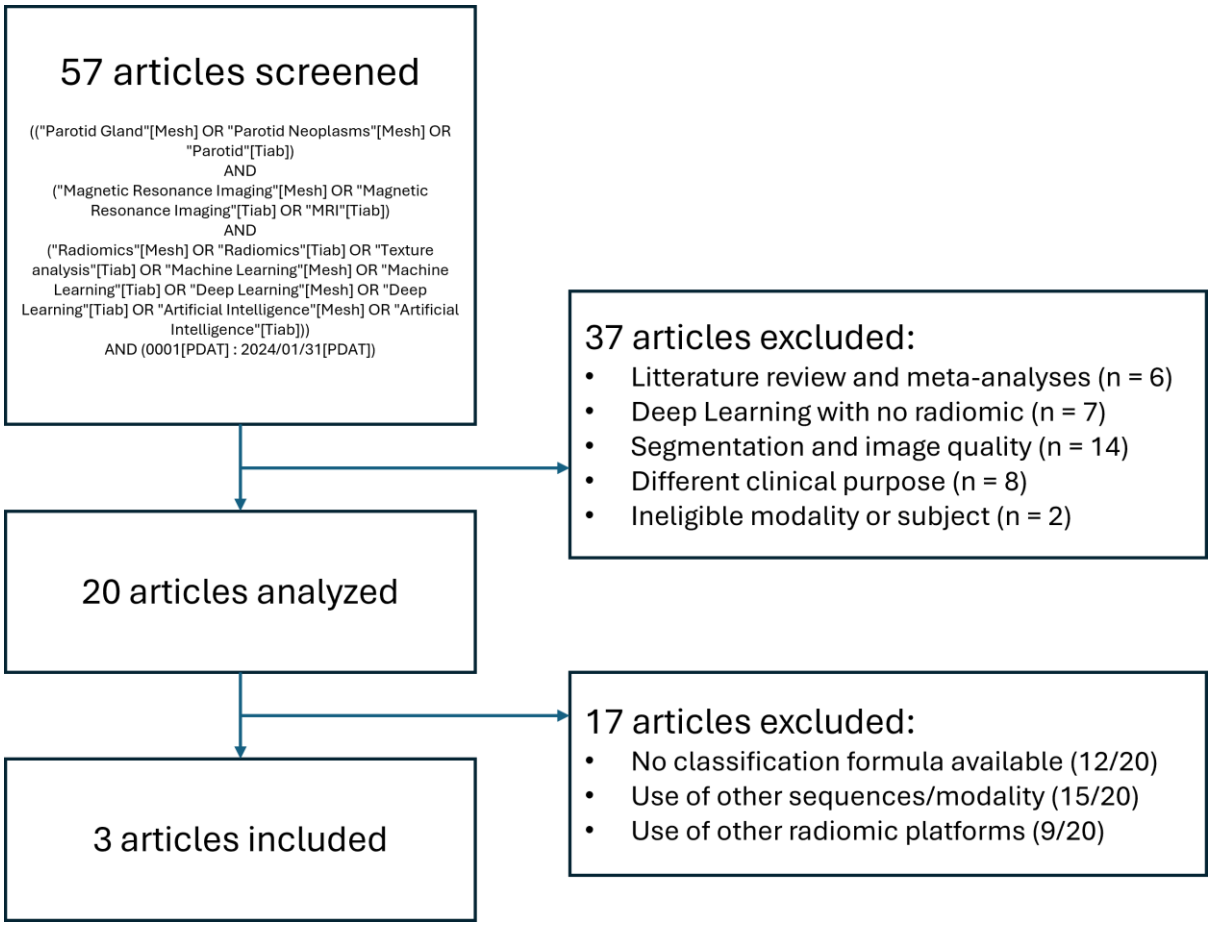
The selection process resulted in the selection of 3 articles

- Article 1 : Zheng Y, Li J, Liu S, et al (2020) MRI-Based radiomics nomogram for differentiation of benign and malignant lesions of the parotid gland. Eur Radiol. https://doi.org/10.1007/s00330-020-07483-4

- Article 2 : Zheng Y, Chen J, Xu Q, et al (2021) Development and validation of an MRI-based radiomics nomogram for distinguishing Warthin's tumour from pleomorphic adenomas of the parotid gland. Dentomaxillofac Radiol 50:20210023. https://doi.org/10.1259/dmfr.20210023

- Article 3 : Hu Z, Guo J, Feng J, et al (2022) Value of T2-weighted-based radiomics model in distinguishing Warthin tumor from pleomorphic adenoma of the parotid. Eur Radiol 33:4453–4463. https://doi.org/10.1007/s00330-022-09295-0

-



Supp Figure 1: Diagram of step-by-step exclusion of studies, from the initial identification in PubMed to the final set of articles included

First author and reference	Year	Clinical question	Training population (n)	Test population (n)	MRI sequences	Radiomic platform	pre treatment and radiomic setting description completeness	Classifier	classification formula availability	Validation method	Performance summary
Yang [2]	2024	PMA vs WT	88	38 (internal test) 23 (external test)	T2, T1, FS-T1 CE	PyRadiomics	partial	logistic regression	1	Internal and external test set	AUC external validation 0.915
Muntean [3]	2023	MT vs BT	83	25	T2, CE-FS-T1	PyRadiomics	Yes	LASSO regression	1	Internal test set	AUC test 0.786
Muraoka [4]	2023	PMA vs WT	22	None	STIR, ADC	MaZda	Yes	"combination"	0	None	AUC 0.93 – 0.96
Fathi Kazerooni [5]	2022	MT vs BT	31	None	T2, ADC, DCE-T1	In house software	No	SVM	0	None	Accuracy 1.00

Hu [6]	2023	PMA vs WT	82	35	T2	PyRadiomics	partial	logistic regression	partial	internal test set	AUC validation 0.934
Qi [7]	2022	BT vs MT, PMA vs MT,WT vs MT, PMA vs WT	128	55	FS-T2, ADC, CE-T1	Feature Explorer	no	Linear regression	1	Internal test set	AUC 0.907 (MT vs BT) & 0.967 (PA vs WT)
Faggioni [8]	2022	WT vs PMA	81 (T2), 52 (T1)	None	T2,CE-FS-T1	PyRadiomics	partial	logistic regression	1	None	AUC 0.9 (CE-FS-T1)
He [9]	2022	Classif. 4 subtypes (incl. PMA vs WT)	208	90	T2, T1, CE-FS-T1	PyRadiomics	partial	XGBoost, SVM, Decision Tree	0	Internal test set	XGBoost accuracy 0.71
Wen [10]	2022	BT vs MT, PMA vs WT	91	39	ADC	PyRadiomics	No	linear discriminant analysis	1	Internal test set	AUC 0.76 (BT/MT) & 0.92 (PA/WT)
Juan [11]	2022	MT, PMA, WT	78	N/A	ADC	None (mean and standard deviation)	Yes	random forest	0	Leave one out cross validation	AUC 0.81 (MT)

Vernuccio [12]	2021	BT vs MT, PMA vs WT	57	None	T1, CE-T1 , SPIR, T2	MaZda	No	discriminant analysis	0	5-fold cross validation	AUC 0.927 (BT/MT) & 0.808 (PA/WT)
Piludu [13]	2021	BT vs MT& WT vs MT	69	44	T2, ADC	S IBEX	Yes	SVM	0	External validation test set	Acc 0.87 (WT vs MT) & 0.80 (BT vs MT)
Zheng [14]	2021	WT vs PMA	75	52	T1, FS-T2	PyRadiomics	partial	Logistic regression	1	External validation test set	AUC validation 0.918
Song [15]	2021	WT vs PMA	126	76	T1, T2	IBEX	no	multivariable logistic regression, SVM	0	Internal validation test set	AUC validation 0.90
Liu [16]	2021	WT vs PMA	626	N/A	T1, T2 and CT	MaZda	partial	LASSO regression	1	None	AUC MRI 0.911
Nardi [17]	2021	MT vs BT, PMA vs WT vs EM vs Ly	54	None	ADC	LIFEx	partial	Selected features evaluated separately	1	None	AUC 0.81 MT vs BT

Liu [18]	2021	MT vs BT	74	35	T1, T2	MaZda	No	multivariable logistic regression	1	External validation test set	AUC 0.76 on validation cohort
Zheng [19]	2021	MT vs BT	80	35	T1, FS-T2	PyRadiomics	partial	LASSO regression	1	External validation test set	AUC validation 0.938
Sarioglu [20]	2020	MT vs BT	95	N/A	FS-T2, CE-T1	LIFEx	partial	Selected features evaluated separately	1	None	Spe 0.99 Sen 0.50
Fruehwald-Pallamar [21]	2013	BT vs MT, PMA vs WT	38	None	T1,CE-T1, STIR, ADC	MaZda	partial	linear discriminant analysis with k nearest neighbor classification	1	None	Max accuracy 0.83 (MT vs BT)

Supp Table 1: Details of the 20 analyzed articles. The lines of the 3 included articles are shaded.
BT: benign tumor, MT: malignant tumor, PMA: Pleiomorphic adenoma, WT: Whartin's tumor, EM: Epithelial Malignancy, Ly: Lymphoma
FS: Fat Saturation, CE: Contrast Enhanced, DCE : Dynamic Contrast Enhanced

REFERENCES :

1. Park JE, Kim D, Kim HS, et al (2020) Quality of science and reporting of radiomics in oncologic studies: room for improvement according to radiomics quality score and TRIPOD statement. Eur Radiol 30:523–536. https://doi.org/10.1007/s00330-019-06360-z

2. Yang J, Bi Q, Jin Y, et al (2024) Different MRI-based radiomics models for differentiating misdiagnosed or ambiguous pleomorphic adenoma and Warthin tumor of the parotid gland: a multicenter study. Front Oncol 14:1392343. https://doi.org/10.3389/fonc.2024.1392343

3. Muntean DD, Dudea SM, Băciuț M, et al (2023) The Role of an MRI-Based Radiomic Signature in Predicting Malignancy of Parotid Gland Tumors. Cancers 15:3319. https://doi.org/10.3390/cancers15133319

4. Muraoka H, Kaneda T, Kondo T, et al (2023) Differential diagnosis of parotid gland tumors using apparent diffusion coefficient, texture features, and their combination. Dentomaxillofac Radiol 52:20220404. https://doi.org/10.1259/dmfr.20220404

5. Fathi Kazerooni A, Nabil M, Alviri M, et al (2022) Radiomic Analysis of Multi-parametric MR Images (MRI) for Classification of Parotid Tumors. J Biomed Phys Eng 12:599–610. https://doi.org/10.31661/jbpe.v0i0.2007-1140

6. Hu Z, Guo J, Feng J, et al (2022) Value of T2-weighted-based radiomics model in distinguishing Warthin tumor from pleomorphic adenoma of the parotid. Eur Radiol 33:4453–4463. https://doi.org/10.1007/s00330-022-09295-0

7. Qi J, Gao A, Ma X, et al (2022) Differentiation of Benign From Malignant Parotid Gland Tumors Using Conventional MRI Based on Radiomics Nomogram. Front Oncol 12:937050. https://doi.org/10.3389/fonc.2022.937050

8. Faggioni L, Gabelloni M, De Vietro F, et al (2022) Usefulness of MRI-based radiomic features for distinguishing Warthin tumor from pleomorphic adenoma: performance assessment using T2-weighted and post-contrast T1-weighted MR images. Eur J Radiol Open 9:100429. https://doi.org/10.1016/j.ejro.2022.100429

9. He Z, Mao Y, Lu S, et al (2022) Machine learning–based radiomics for histological classification of parotid tumors using morphological MRI: a comparative study. Eur Radiol. https://doi.org/10.1007/s00330-022-08943-9

10. Wen B, Zhang Z, Zhu J, et al (2022) Apparent Diffusion Coefficient Map–Based Radiomics Features for Differential Diagnosis of Pleomorphic Adenomas and Warthin Tumors From Malignant Tumors. Front Oncol 12:830496. https://doi.org/10.3389/fonc.2022.830496

11. Juan C, Huang T, Liu Y, et al (2022) Improving diagnosing performance for malignant parotid gland tumors using machine learning with multifeatures based on diffusion-weighted magnetic resonance imaging. NMR in Biomedicine 35:e4642. https://doi.org/10.1002/nbm.4642

12. Vernuccio F, Arnone F, Cannella R, et al (2021) Diagnostic performance of qualitative and radiomics approach to parotid gland tumors: which is the added benefit of texture analysis? BJR 20210340. https://doi.org/10.1259/bjr.20210340

13. Piludu F, Marzi S, Ravanelli M, et al (2021) MRI-Based Radiomics to Differentiate between Benign and Malignant Parotid Tumors With External Validation. Front Oncol 11:656918. https://doi.org/10.3389/fonc.2021.656918

14. Zheng Y, Chen J, Xu Q, et al (2021) Development and validation of an MRI-based radiomics nomogram for distinguishing Warthin's tumour from pleomorphic adenomas of the parotid gland. Dentomaxillofac Radiol 50:20210023. https://doi.org/10.1259/dmfr.20210023

15. Song X (2021) Radiomics derived from dynamic contrast-enhanced MRI pharmacokinetic protocol features: the value of precision diagnosis ovarian neoplasms. Eur Radiol 11

16. Liu Y, Zheng J, Lu X, et al (2021) Radiomics-based comparison of MRI and CT for differentiating pleomorphic adenomas and Warthin tumors of the parotid gland: a retrospective study. Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology 131:591–599. https://doi.org/10.1016/j.oooo.2021.01.014

17. Nardi C, Tomei M, Pietragalla M, et al (2021) Texture analysis in the characterization of parotid salivary gland lesions: A study on MR diffusion weighted imaging. European Journal of Radiology 136:109529. https://doi.org/10.1016/j.ejrad.2021.109529

18. Liu Y, Zheng J, Zhao J, et al (2021) Magnetic resonance image biomarkers improve differentiation of benign and malignant parotid tumors through diagnostic model analysis. Oral Radiol 37:658–668. https://doi.org/10.1007/s11282-020-00504-4

19. Zheng Y, Li J, Liu S, et al (2020) MRI-Based radiomics nomogram for differentiation of benign and malignant lesions of the parotid gland. Eur Radiol. https://doi.org/10.1007/s00330-020-07483-4

20. Sarioglu O, Sarioglu FC, Akdogan AI, et al (2020) MRI-based texture analysis to differentiate the most common parotid tumours. Clinical Radiology 75:877.e15-877.e23. https://doi.org/10.1016/j.crad.2020.06.018

21. Fruehwald-Pallamar J, Czerny C, Holzer-Fruehwald L, et al (2013) Texture-based and diffusion-weighted discrimination of parotid gland lesions on MR images at 3.0 Tesla. NMR Biomed 26:1372–1379. https://doi.org/10.1002/nbm.2962

2 Electronic Supplementary Material 2

Details of the models for article 1

Formula of the Radscore for article 1

Radscore = 8.131645 - 0.02744869* T1_original_firstorder_10Percentile

- 2.670423 * T1_wavelet-HLL_glcm_Idn

- 2.381722* T1_wavelet-LHL_gldm_DependenceEntropy

- 1.750878 * T1_wavelet-LHH_gldm_DependenceVariance

- 0.2131322* T1_wavelet-LHH_firstorder_Energy

- 1.113082e-15 * T1_wavelet-LHH_firstorder_TotalEnergy

+ 0.1038013 * T1_wavelet-HLH_gldm_SmallDependenceLowGrayLevelEmphasis

- 0.220293 * T1_wavelet-HLH_glcm_Correlation

+ 0.1611885 * T1_wavelet-HHL_gldm_SmallDependenceLowGrayLevelEmphasis

- 0.169119 * T1_wavelet-HHL_glcm_Correlation

- 0.1256594 * T1_wavelet-LLL_firstorder_Minimum

- 3.430239 * T2FS_original_glcm_Imc2

+ 0.2339147 * T2FS_wavelet-LHL_firstorder_Mean

+ 0.1476619 * T2FS_wavelet-LHL_ngtdm_Busyness

+ 2.325397 * T2FS_wavelet-HLH_gldm_DependenceEntropy

- 0.36343 * T2FS_wavelet-HLH_glszm_GrayLevelNonUniformityNormalized

+ 0.2407029 * T2FS_wavelet-LLL_firstorder_Kurtosis

Formula of the Nomoscore for article 1

Nomoscore = 0.105 + 4.270 * Radscore + 1.095 * IST + 0.678 * DLI

IST: Infiltration of Surrounding Tissues ; DLI: Deep Lobe Involvement

3 Electronic Supplementary Material 3

Details of the models for article 2

Formula of the Radscore for article 2

Radscore = -1.467 - 0.459 * T1_wavelet-HHL_glcm_Correlation

+ 2.509 * T2FS_wavelet-LLH_glcm_Imc2

- 0.179 * T2FS_original_firstorder_Kurtosis

+ 0.467 * T2FS_wavelet-HLL_firstorder_Kurtosis

+ 0.144 * T2FS_wavelet-LLH_glcm_Correlation

- 0.131 * T2FS_wavelet-LLH_glszm_SizeZoneNonUniformityNormalized

+ 0.029 * T2FS_wavelet-HLH_glcm_ClusterShade

- 0.075 * T2FS_wavelet-HHL_glcm_ClusterShade

+ 0.067 * T2FS_wavelet-HHL_glcm_MCC

- 0.136 * T2FS_wavelet-LLL_glcm_ClusterShade

- 0.568 * T2FS_wavelet-LLL_glrlm_GrayLevelNonUniformityNormalized

- 0.534 * T2FS_wavelet-LLL_glszm_GrayLevelNonUniformityNormalized

Formula of the Nomoscore for article 2

Nomoscore = 3.942 + 3.898* Radscore + 0.091* Age

4 Electronic Supplementary Material 4

Details of the models for article 3

Formula of the Radscore for article 3

Radscore = - 2.16328

+ 2.83586 * original_shape_Flatness

+ 0.00117 * original_firstorder_10Percentile

- 4.90062 * wavelet-LLH_glcm_Imc1

- 0.92691 * wavelet-LLH_gldm_LowGrayLevelEmphasis

+ 0.00133 * wavelet-LLL_firstorder_10Percentile

- 0.02628 * wavelet-LLL_gldm_DependenceVariance

- 0.00449 * wavelet-LLL_gldm_LargeDependenceEmphasis

Formula of the Nomoscore for article 3

Points = 141.6811 + 15.29 * radscore + 1.111111 * age + 17.7 * (1-parotid tail)

Risk (Nomoscore) = 0.02781 * Points - 2.01682

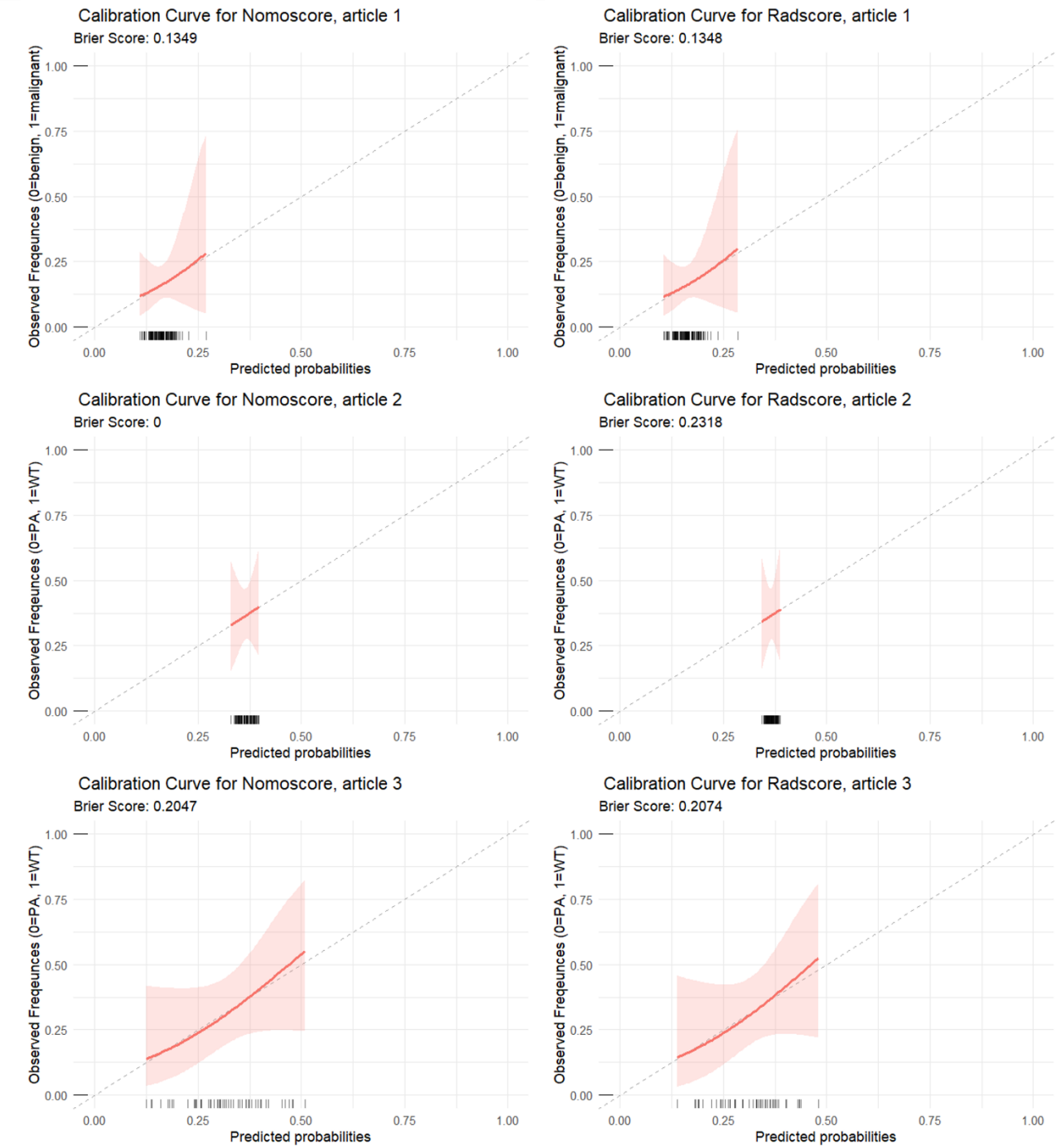
5 Electronic Supplementary Material 5

Distribution of histologic types for article 1

Malignant lesions	n=21	Benign lesion	n=109
lymphoma	6	PMA	59
squamous cell carcinoma metastasis	5	WT	34
melanoma metastasis	3	basal cell adenoma	8
squamous cell carcinoma	2	lymphoepithelial cyst	3
Low-grade mucoepidermoid carcinoma	2	oncocytoma	2
Salivary ductal adenocarcinoma	1	hamartoma	1
Acinic cell carcinoma	1	benign myoepithelial lesion	1
carcinoma ex pleomorphic adenoma	1	lymphadenoma	1
poorly differentiated carcinoma	1		
primary papillary cystadenocarcinoma	1		

6 Electronic Supplementary Material 6

Brier scores and calibration curves for all 6 models



7 Electronic Supplementary Material 7

Evaluation of the radiomic quality scores for all 3 studies

RQS for studies 1 and 2

Image protocol quality - well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability

- [x] protocols well documented
- [ ] public protocol used
- [ ] none

Multiple segmentations - possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyse feature robustness to segmentation variabilities

- (•) yes
- ( ) no

Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyse feature robustness to these sources of variability

- ( ) yes
- (•) no

Imaging at multiple time points • collect images of individuals at additional time points. Analyse feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage)

- ( ) yes
- (•) no

Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features

- (•) Either measure is implemented
- ( ) Neither measure is implemented

Multivariable analysis with non radiomics features (for example, EGFR mutation) - is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non radiomics features

- (•) yes
- ( ) no

Detect and discuss biological correlates - demonstration of phenotypic differences (possibly associated with underlying gene–protein expression patterns) deepens understanding of radiomics and biology

- ( ) yes
- (•) no

Cut-off analyses • determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results

- ○ yes
- ◉ no

Discrimination statistics - report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, p-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)

- ☑ a discrimination statistic and its statistical significance are reported
- ☐ a resampling method technique is also applied
- ☐ none

Calibration statistics - report calibration statistics (for example, Calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, P-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)

- ☑ a calibration statistic and its statistical significance are reported
- ☐ a resampling method technique is applied
- ☐ none

Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker

- ○ yes
- ◉ no

Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance

- ☐ No validation
- ☐ validation is based on a dataset from the same institute
- ☑ validation is based on a dataset from another institute
- ☐ validation is based on two datasets from two distinct institutes
- ☐ the study validates a previously published signature
- ☐ validation is based on three or more datasets from distinct institutes

Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics

- ◉ yes
- ○ no

Potential clinical utility - report on the current and potential application of the model in a clinical setting (for example, decision curve analysis).

- ○ yes
- ◉ no

Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (for example, QALYs generated)

- ○ yes
- ◉ no

Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study

- ☐ scans are open source
- ☐ region of interest segmentations are open source
- ☐ the code is open sourced
- ☐ radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source

Total score **13** (36.11%)

METRICS for studies 1 and 2

Items/Conditions	Definitions	Weights	
Study Design			
Item#1	? Adherence to radiomics and/or machine learning-specific checklists or guidelines	0.0368	no
Item#2	? Eligibility criteria that describe a representative study population	0.0735	yes
Item#3	? High-quality reference standard with a clear definition	0.0919	yes
Imaging Data			
Item#4	? Multi-center	0.0438	yes
Item#5	? Clinical translatability of the imaging data source for radiomics analysis	0.0292	yes
Item#6	? Imaging protocol with acquisition parameters	0.0438	yes
Item#7	? The interval between imaging used and reference standard	0.0292	yes
SegmentationC			
Condition#1	? Does the study include segmentation?		yes
Condition#2	? Does the study include fully automated segmentation?		no
Item#8	? Transparent description of segmentation methodology	0.0337	no
Item#9	? Formal evaluation of fully automated segmentationC	0.0225	n/a
Item#10	? Test set segmentation masks produced by a single reader or automated tool	0.0112	yes
Image Processing and Feature Extraction			
Condition#3	? Does the study include hand-crafted feature extraction?		yes
Item#11	? Appropriate use of image preprocessing techniques with transparent description	0.0622	yes
Item#12	? Use of standardized feature extraction softwareC	0.0311	yes
Item#13	? Transparent reporting of feature extraction parameters, otherwise providing a default configuration statement	0.0415	no
Feature Processing			
Condition#4	? Does the study include tabular data?		yes
Condition#5	? Does the study include end-to-end deep learning?		no
Item#14	? Removal of non-robust featuresC	0.0200	yes
Item#15	? Removal of redundant featuresC	0.0200	yes
Item#16	? Appropriateness of dimensionality compared to data sizeC	0.0300	yes
Item#17	? Robustness assessment of end-to-end deep learning pipelinesC	0.0200	n/a
Preparation for Modeling			
Item#18	? Proper data partitioning process	0.0599	yes
Item#19	? Handling of confounding factors	0.0300	no
Metrics and Comparison			

Item#20	? Use of appropriate performance evaluation metrics for task	0.0352	no
Item#21	? Consideration of uncertainty	0.0234	yes
Item#22	? Calibration assessment	0.0176	yes
Item#23	? Use of uni-parametric imaging or proof of its inferiority	0.0117	no
Item#24	? Comparison with a non-radiomic approach or proof of added clinical value	0.0293	no
Item#25	? Comparison with simple or classical statistical models	0.0176	no
Testing			
Item#26	? Internal testing	0.0375	yes
Item#27	? External testing	0.0749	yes
Open Science			
Item#28	? Data availability	0.0075	no
Item#29	? Code availability	0.0075	no
Item#30	? Model availability	0.0075	no
Total METRICS score:			**73.0%**
? Quality category:			**Good**

RQS for study 3

Image protocol quality • well-documented image protocols (for example, contrast, slice thickness, energy, etc.) and/or usage of public image protocols allow reproducibility/replicability

- [x] protocols well documented
- [ ] public protocol used
- [ ] none

Multiple segmentations - possible actions are: segmentation by different physicians/algorithms/software, perturbing segmentations by (random) noise, segmentation at different breathing cycles. Analyse feature robustness to segmentation variabilities

- (•) yes
- ( ) no

Phantom study on all scanners - detect inter-scanner differences and vendor-dependent features. Analyse feature robustness to these sources of variability

- ( ) yes
- (•) no

Imaging at multiple time points - collect images of individuals at additional time points. Analyse feature robustness to temporal variabilities (for example, organ movement, organ expansion/shrinkage)

- ( ) yes
- (•) no

Feature reduction or adjustment for multiple testing - decreases the risk of overfitting. Overfitting is inevitable if the number of features exceeds the number of samples. Consider feature robustness when selecting features

- (•) Either measure is implemented
- ( ) Neither measure is implemented

Multivariable analysis with non-radiomics features (for example, EGFR mutation) • is expected to provide a more holistic model. Permits correlating/inferencing between radiomics and non-radiomics features

- (•) yes
- ( ) no

Detect and discuss biological correlates • demonstration of phenotypic differences (possibly associated with underlying gene–protein expression patterns) deepens understanding of radiomics and biology

- ( ) yes
- (•) no

Cut-off analyses - determine risk groups by either the median, a previously published cut-off or report a continuous risk variable. Reduces the risk of reporting overly optimistic results

- ○ yes
- ◉ no

Discrimination statistics - report discrimination statistics (for example, C-statistic, ROC curve, AUC) and their statistical significance (for example, p-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)

- ☑ a discrimination statistic and its statistical significance are reported
- ☐ a resampling method technique is also applied
- ☐ none

Calibration statistics - report calibration statistics (for example, Calibration-in-the-large/slope, calibration plots) and their statistical significance (for example, P-values, confidence intervals). One can also apply resampling method (for example, bootstrapping, cross-validation)

- ☑ a calibration statistic and its statistical significance are reported
- ☐ a resampling method technique is applied
- ☐ none

Prospective study registered in a trial database - provides the highest level of evidence supporting the clinical validity and usefulness of the radiomics biomarker

- ○ yes
- ◉ no

Validation - the validation is performed without retraining and without adaptation of the cut-off value, provides crucial information with regard to credible clinical performance

- ☐ No validation
- ☑ validation is based on a dataset from the same institute
- ☐ validation is based on a dataset from another institute
- ☐ validation is based on two datasets from two distinct institutes
- ☐ the study validates a previously published signature
- ☐ validation is based on three or more datasets from distinct institutes

Comparison to 'gold standard' - assess the extent to which the model agrees with/is superior to the current 'gold standard' method (for example, TNM-staging for survival prediction). This comparison shows the added value of radiomics

- ◉ yes
- ○ no

Potential clinical utility - report on the current and potential application of the model in a clinical setting (for example, decision curve analysis).

- ○ yes
- ◉ no

Cost-effectiveness analysis - report on the cost-effectiveness of the clinical application (for example, QALYs generated)

- ○ yes
- ◉ no

Open science and data - make code and data publicly available. Open science facilitates knowledge transfer and reproducibility of the study

- ☐ scans are open source
- ☐ region of interest segmentations are open source
- ☐ the code is open soured
- ☐ radiomics features are calculated on a set of representative ROIs and the calculated features and representative ROIs are open source

Total score **12** (33.33%)

METRICS for study 3

Items/Conditions	Definitions	Weights	
Study Design			
Item#1	? Adherence to radiomics and/or machine learning-specific checklists or guidelines	0.0368	no
Item#2	? Eligibility criteria that describe a representative study population	0.0735	yes
Item#3	? High-quality reference standard with a clear definition	0.0919	yes
Imaging Data			
Item#4	? Multi-center	0.0438	no
Item#5	? Clinical translatability of the imaging data source for radiomics analysis	0.0292	yes
Item#6	? Imaging protocol with acquisition parameters	0.0438	yes
Item#7	? The interval between imaging used and reference standard	0.0292	yes
SegmentationC			
Condition#1	? Does the study include segmentation?		yes
Condition#2	? Does the study include fully automated segmentation?		no
Item#8	? Transparent description of segmentation methodology	0.0337	no
Item#9	? Formal evaluation of fully automated segmentationC	0.0225	n/a
Item#10	? Test set segmentation masks produced by a single reader or automated tool	0.0112	yes
Image Processing and Feature Extraction			
Condition#3	? Does the study include hand-crafted feature extraction?		yes
Item#11	? Appropriate use of image preprocessing techniques with transparent description	0.0622	no
Item#12	? Use of standardized feature extraction softwareC	0.0311	yes
Item#13	? Transparent reporting of feature extraction parameters, otherwise providing a default configuration statement	0.0415	no
Feature Processing			
Condition#4	? Does the study include tabular data?		yes
Condition#5	? Does the study include end-to-end deep learning?		no
Item#14	? Removal of non-robust featuresC	0.0200	yes
Item#15	? Removal of redundant featuresC	0.0200	yes
Item#16	? Appropriateness of dimensionality compared to data sizeC	0.0300	yes
Item#17	? Robustness assessment of end-to-end deep learning pipelinesC	0.0200	n/a
Preparation for Modeling			
Item#18	? Proper data partitioning process	0.0599	yes
Item#19	? Handling of confounding factors	0.0300	no

Metrics and Comparison			
Item#20	? Use of appropriate performance evaluation metrics for task	0.0352	no
Item#21	? Consideration of uncertainty	0.0234	yes
Item#22	? Calibration assessment	0.0176	yes
Item#23	? Use of uni-parametric imaging or proof of its inferiority	0.0117	no
Item#24	? Comparison with a non-radiomic approach or proof of added clinical value	0.0293	no
Item#25	? Comparison with simple or classical statistical models	0.0176	no
Testing			
Item#26	? Internal testing	0.0375	yes
Item#27	? External testing	0.0749	no
Open Science			
Item#28	? Data availability	0.0075	no
Item#29	? Code availability	0.0075	no
Item#30	? Model availability	0.0075	no
Total METRICS score:			**54.1%**
? Quality category:			**Moderate**