DIR

Diagnostic and Interventional
Radiology

Contents

Full text of these articles can be accessed online at www.dirjournal.org or through PubMed (https://www.ncbi.nlm.nih.gov/pmc/journals/2754/).

Long-term follow-up results of multiparametric prostate MRI and the prognostic value of PI-RADS: a single-center retrospective cohort study

- Ömer Önder[1]
- Müjdat Ayva[2]
- Yasin Yaraşır[1]
- Volkan Gürler[1]
- Mustafa Sertaç Yazıcı[2]
- Bülent Akdoğan[2]
- Ali Devrim Karaosmanoğlu[1]
- Muşturay Karçaaltıncaba[1]
- Mustafa Nasuh Özmen[1]
- Deniz Akata[1]

[1]Hacettepe University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

[2]Hacettepe University Faculty of Medicine, Department of Urology, Ankara, Türkiye

PURPOSE

We aim to examine the long-term outcomes of patients who underwent multiparametric prostate magnetic resonance imaging (mp-MRI) for suspected prostate cancer (PCa), specifically based on their initial Prostate Imaging Reporting and Data System (PI-RADS) categories and various clinical factors. Our secondary aim is to evaluate the prognostic value of the PI-RADS through the National Comprehensive Cancer Network (NCCN) risk group distribution.

METHODS

This research was conducted as a single-center retrospective cohort study in a tertiary care hospital. A total of 1,359 cases having at least one histopathological examination after the initial mp-MRI and/or adequate clinical/radiological follow-up data were included in the clinically significant PCa (cs-PCa) diagnosis-free survival analysis. Initial mp-MRI dates were accepted as the start of follow-up for the time-to-event analysis. The event was defined as cs-PCa diagnosis (International Society of Urological Pathology ≥2). Patients who were not diagnosed with cs-PCa during follow-up were censored according to predefined literature-based criteria at the end of the maximum follow-up duration with no reasonable suspicion of PCa and no biopsy indication. The impact of various factors on survival was assessed using a log-rank test and multivariable Cox regression. Subsequently, 394 cases diagnosed with PCa during follow-up were evaluated, based on initial PI-RADS categories and NCCN risk groups.

RESULTS

Three main risk factors for cs-PCa diagnosis during follow-up were an initial PI-RADS 5 category, initial PI-RADS 4 category, and high MRI-defined PSA density (mPSAD), with average hazard ratios of 29.52, 14.46, and 3.12, respectively. The PI-RADS 3 category, advanced age group, and biopsy-naïve status were identified as additional risk factors (hazard ratios: 2.03, 1.54–1.98, and 1.79, respectively). In the PI-RADS 1–2 cohort, 1, 3, and 5-year cs-PCa diagnosis-free survival rates were 99.1%, 96.5%, and 93.8%, respectively. For the PI-RADS 3 cohort, 1, 3, and 5-year cs-PCa diagnosis-free survival rates were 94.9%, 90.9%, and 89.1%, respectively. For the PI-RADS 4 cohort, 1, 3, and 5-year cs-PCa diagnosis-free survival rates were 56.6%, 55.1%, and 55.1%, respectively. These rates were found to all be 24.2% in the PI-RADS 5 cohort. Considering the 394 cases diagnosed with PCa during follow-up, PI-RADS ≥4 cases were more likely to harbor unfavorable PCa compared to PI-RADS ≤3 cases ($P < 0.001$). In the PI-RADS 3 subgroup analysis, a low mPSAD (<0.15 ng/mL^2) was found to be a protective prognostic factor against unfavorable PCa ($P = 0.005$).

CONCLUSION

The PI-RADS category has a significant impact on patient management and provides important diagnostic and prognostic information. Higher initial PI-RADS categories are associated with decreased follow-up losses, a shorter time to PCa diagnosis, increased biopsy rates, a higher likelihood of developing cs-PCa during follow-up, and a worse PCa prognosis. Combining mPSAD with PI-RADS categories could enhance diagnostic stratification in the identification of cs-PCa.

KEYWORDS

Prostatic neoplasm, follow-up studies, multiparametric magnetic resonance imaging, prognosis, diagnosis, biopsy

Corresponding author: Deniz Akata

E-mail: akataden@gmail.com

Prostate cancer (PCa) is one of the most common cancers among men worldwide. It encompasses a broad and heterogeneous spectrum of diseases, ranging from low-grade clinically insignificant tumors to metastatic disease with high morbidity and mortality. The clinical heterogeneity and significant differences in prognosis have led to the search for risk stratification in PCa management. Classification systems incorporating biochemical, clinical, and histopathological data, such as the National Comprehensive Cancer Network (NCCN) guidelines and the D'Amico risk scale, are frequently used in urology practice and effectively employed in patient management.[1,2]

As another effective tool, multiparametric prostate magnetic resonance imaging (mp-MRI) is widely used for the early detection, staging, and monitoring of prostate tumors. Efforts to integrate mp-MRI with diagnostic algorithms have gained momentum following the implementation of the Prostate Imaging Reporting and Data System (PI-RADS) categories. This standardized reporting system has facilitated the assessment of mp-MRI findings and enabled objective stratification for suspected clinically significant PCa (cs-PCa).[3] The increasing body of literature elucidated the potential diagnostic value of mp-MRI, and, eventually, the European Association of Urology guideline recommended conducting an mp-MRI before the initial biopsy.[4]

Our study aims to evaluate the long-term follow-up outcomes of patients who underwent mp-MRI for suspected PCa, explicitly focusing on initial PI-RADS assessment categories. We will determine cs-PCa diagnosis-free survival probabilities across different PI-RADS cohorts, calculate the hazard ratios of key clinical parameters influencing the outcome, assess the prognostic value of the PI-RADS by examining NCCN risk group distribution, and discuss possible management strategies for different patient subgroups in light of our results and current literature.

Methods

This study was approved by the Hacettepe University Non-Invasive Clinical Research Ethics Committee (decision number: 2022/01-30, date: 10.05.2022), with a waiver of informed consent.

The workflow of the study is summarized in Figure 1.

Data collection

Patients who underwent mp-MRI in our institution between April 2014 and June 2021 were identified using the hospital's information system. Relevant clinical, radiological, and histopathological data were extracted for these patients throughout their follow-up period (until June 2022).

Definitions and basic considerations

Baseline prostate-specific antigen (PSA): PSA values at the time of initial mp-MRI were considered as the baseline PSA.

Prostate volume: Prostate volume was calculated using the ellipsoid formula (length × width × height × π/6) on the initial mp-MRI.[5]

MRI-defined PSA density (mPSAD): mPSAD was calculated by dividing the baseline PSA by the prostate volume. A cut-off value of 0.15 ng/mL^2 was used for mPSAD.[4]

PSA velocity (PSAV): PSAV was calculated using the first-to-last method in patients with at least three PSA values covering at least one year of follow-up interval.[6]

mp-MRI evaluation: The PI-RADS categories of cases were extracted directly from the mp-MRI reports in the hospital's information system and were used without making any retrospective category changes, even in the presence of radiopathology discrepancies.

This methodology aimed to maintain the cause-and-effect relationship between the already reported PI-RADS categories and the subsequent diagnostic management, ensure that category assigners remained blinded and unbiased, and obtain results that reflected everyday practice rather than ideal conditions. In accordance with our center's routine radiology practices, each mp-MRI was categorized into one of four PI-RADS assessment groups (PI-RADS 1–2, 3, 4, and 5 categories). These assignments were carried out by one of four readers, each with at least 10 years of experience in abdominal radiology and following the guidelines of PI-RADSv2 and PI-RADSv2.1. All mp-MRI examinations were performed on five MRI (two 3.0 Tesla and three 1.5 Tesla) scanners, using imaging protocols in line with PI-RADS recommendations. The imaging protocols of the two most commonly used devices are provided as Supplementary Table 1.

Histopathological examination types: In our institution, PI-RADS 1–2 cases requiring biopsy undergo transrectal ultrasound (TRUS)-guided systematic biopsy in the urology department and rarely in the interventional radiology unit. On the other hand, PI-RADS ≥3 cases typically undergo MRI-TRUS fusion biopsy combined with systematic biopsy, performed by interventional radiologists.

Histopathological examination results: All available core biopsy results and, if present, radical prostatectomy results of the patients were extracted from the hospital's information system. Normal prostate tissue, inflammation, atypical small acinar proliferation, and prostatic intraepithelial neoplasia were considered non-neoplastic. In cases diagnosed with PCa, biopsy results were recorded according to the International Society of Urological Pathology (ISUP) grade group system. The highest ISUP grade in the sample was accepted as the final result; ISUP = 1 PCa cases were classified as clinically insignificant PCa, and ISUP ≥2 cases were considered cs-PCa.

NCCN risk groups: The risk stratification of patients was completed by urologists using the clinical and histopathological results according to the NCCN guidelines. Patients included in the very low, low, and intermediate-favorable risk groups, which had the option of undergoing active surveillance, were accepted as "favorable PCa". The intermediate-unfavorable, high, and very high-risk groups were categorized as "unfavorable PCa".[1]

Figure 1. Workflow of the study. PCa, prostate cancer; PSA, prostate-specific antigen; mp-MRI, multiparametric prostate magnetic resonance imaging; PI-RADS, Prostate Imaging Reporting and Data System; cs-PCa, clinically significant prostate cancer; NCCN, National Comprehensive Cancer Network.

Table 1. Descriptive statistics of the cases included in the survival analysis

Variables		PI-RADS 1–2 cohort (n = 385)	PI-RADS 3 cohort (n = 478)	PI-RADS 4–5 cohort (n = 496)
Age (years)	Mean ± SD	61.55 ± 7.67	61.36 ± 7.70	64.36 ± 8.16
	Median (min–max)	61.00 (37.00–86.00)	61.50 (38.00–85.00)	64.00 (41.00–85.00)
Baseline total PSA (ng/mL)	Mean ± SD	6.83 ± 5.40	7.80 ± 5.13	19.55 ± 102.97
	Median (min–max)	5.66 (0.19–69.64)	5.48 (0.27–41.43)	7.81 (0.34–1968.56)
MRI-defined PSA density (ng/mL^2)	Mean ± SD	0.11 ± 0.09	0.13 ± 0.10	0.47 ± 3.37
	Median (min–max)	0.09 (0.01–0.75)	0.10 (0.01-0.84)	0.15 (0.00–61.86)
Prior negative biopsy, n (%)	Present	117 (30.4%)	156 (32.6%)	128 (25.8%)
	Absent (biopsy-naïve)	268 (69.6%)	322 (67.4%)	368 (74.2%)
Magnetic field strength of the initial MRI scanner, n (%)	1.5 Tesla	181 (47.0%)	183 (38.3%)	174 (35.1%)
	3.0 Tesla	204 (53.0%)	295 (61.7%)	322 (64.9%)
Cases with at least one follow-up mp-MRI, n (%)		153 (39.7%)	177 (37.0%)	72 (14.5%)
Cases with at least one histopathological examination during follow-up, n (%)		101 (26.2%)	240 (50.2%)	472 (95.2%)

PI-RADS, Prostate Imaging Reporting and Data System; MRI, magnetic resonance imaging; PSA, prostate-specific antigen; mp-MRI, multiparametric prostate MRI; SD, standard deviation; min, minimum; max, maximum.

Case selection: The main dataset was created by excluding patients with no suspicion of PCa, a previous diagnosis of PCa, a history of radical prostatectomy, an unknown baseline total PSA value, and low-quality initial mp-MRI precluding PI-RADS category assessment. Subsequently, patients with at least one histopathological examination and/or at least one follow-up mp-MRI performed a minimum of 6–12 months apart, and/or clinical follow-up records spanning at least 1 year in the urology clinic, were included in the cs-PCa diagnosis-free survival analysis. A small number of indeterminate cases (n = 25) that did not show any stable course and had an unknown outcome were excluded from the survival analysis (Figure 1).

Follow-up considerations

Start of follow-up: Defined as the date of the first mp-MRI.

Event: Cs-PCa (ISUP ≥2) diagnosis based on histopathological examination.

PI-RADS cohort: A group of patients who shared the same initial PI-RADS category and were eligible for cs-PCa diagnosis-free survival analysis.

Censored observations: Patients not diagnosed with cs-PCa during the follow-up were censored based on our pre-established censoring criteria. This censoring was performed at the end of the maximum follow-up period without suspicion of malignancy and without the need for biopsy. Censoring points were decided according to the histopathological examination results, radiological follow-up findings, and clinical stability during follow-up (in order of decreasing significance) to determine cs-PCa diagnosis-free survival times.

Censoring based on histopathological examination: Non-neoplastic or ISUP = 1 PCa results.

Censoring based on follow-up mp-MRI: PI-RADS 1–2, stable or regressing PI-RADS 3, a PI-RADS downgrade from PI-RADS 4–5 to PI-RADS 3, and a PI-RADS upgrade from PI-RADS 1–2 to PI-RADS 3 with at least 1-year subsequent clinical stability and/or subsequent radiological stability.

Censoring based on clinical follow-up

Cases meeting the following three criteria were censored based on clinical follow-up:

i) Final total PSA value not exceeding 20 ng/mL,

ii) Total PSA value within age-based normal range or PSAV below 0.75 ng/mL/year,[7,8]

iii) No suspicious digital rectal examination (DRE) findings

Our schematic algorithm and illustrative case examples explaining follow-up evaluation in detail are given as Supplementary Figures 1 and 2, respectively.

Statistical analysis

The SPSS Statistics (v.11.5) software (Chicago, SPSS Inc.) was used for conducting data analysis. Descriptive statistics were presented in the form of mean ± standard deviation or standard error, median (minimum–maximum) for quantitative variables, and number of patients (percentage) for qualitative variables. A chi-square test was used to examine the association between two categorical variables. An independent samples t-test and one-way analysis of variance were utilized to compare the means of independent groups. A Kaplan–Meier test was used for survival analysis. The impact of various factors on survival was assessed using a log-rank test and multivariable Cox regression analysis. The P value threshold for statistical significance was accepted as 0.05.

Results

Basic descriptive statistics of the main dataset by initial PI-RADS categories and the distribution percentages of follow-up status for each PI-RADS subgroup are given in Figure 2. Table 1 presents the descriptive statistics of each PI-RADS cohort eligible for cs-PCa diagnosis-free survival analysis. No statistically significant differences were observed when each PI-RADS subgroup in the main dataset was compared with the corresponding PI-RADS cohort in terms of primary clinical variables ($P > 0.19$). On the other hand, significant differences were identified among PI-RADS 1–2, 3, and 4–5 cohorts regarding baseline PSA and initial mPSAD values, both of which were positively correlated with the PI-RADS categories (P values 0.002 and 0.01, respectively). The PI-RADS 1–2 and 3 cohorts were comparable in terms of age and prior biopsy status (P values 0.72 and 0.48, respectively). However, the PI-RADS 4–5 cohort demonstrated significantly higher age and biopsy-naïve status percentages than the PI-RADS ≤3 cohort (P values <0.001 and 0.02, respectively).

Of the 1,359 cases included in the survival analysis, 252 (18.5%) were diagnosed with cs-PCa at the end of follow-up. The follow-up

PI-RADS 1-2 (n=942)
Median age: 62 (33-86)
Median baseline total PSA: 5.83 ng/mL (0.16-115)
Median MRI-defined PSA density: 0.09 ng/mL² (0.00-2.65)
Biopsy naive patients, n(%): 688 (73.0%)

PI-RADS 3 (n=891)
Median age: 62 (35-89)
Median baseline total PSA: 6.23 ng/mL (0.27-43.36)
Median MRI-defined PSA density: 0.10 ng/mL² (0.01-0.84)
Biopsy naive patients, n(%): 627 (70.4%)

PI-RADS 4-5 (n=623)
Median age: 64 (41-87)
Median baseline total PSA: 7.69 ng/mL (0.34-1968.56)
Median MRI-defined PSA density: 0.14 ng/mL² (0.00-61.86)
Biopsy naive patients, n(%): 468 (75.1%)

■ Follow-up loss or inadequate follow-up
■ At least one biopsy during follow-up
■ Clinicoradiological follow-up (no biopsy)
■ Only clinical follow-up

Figure 2. In the main dataset, the rates of follow-up loss for Prostate Imaging Reporting and Data System (PI-RADS) 1–2, PI-RADS 3, and PI-RADS 4–5 subgroups are 59%, 46%, and 20%, respectively. Meanwhile, the rates of undergoing at least one biopsy during follow-up are 11%, 27%, and 76% for these categories, in the order given. Consequently, as the initial PI-RADS category increases, follow-up losses decrease, and the probability of undergoing biopsy during follow-up increases. Each distinct PI-RADS subgroup within the main dataset exhibits similar characteristics concerning age, baseline prostate-specific antigen (PSA) levels, magnetic resonance imaging-defined PSA density, and prior biopsy status, with corresponding PI-RADS cohorts included in the survival analysis.
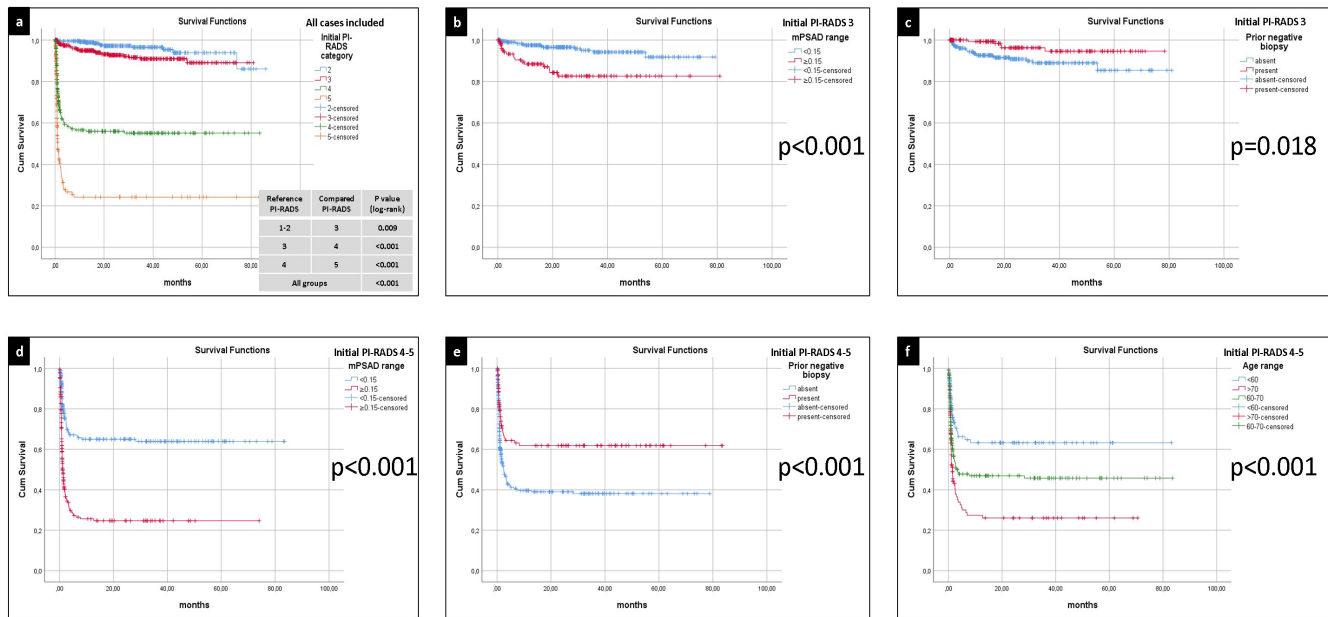
Figure 3. (a) The clinically significant prostate cancer (cs-PCa) diagnosis-free survival curves of all cases included in the survival analysis, stratified by the initial Prostate Imaging Reporting and Data System (PI-RADS) category. Among all PI-RADS cohorts, a statistically significant inverse correlation was identified between the initial PI-RADS category and cs-PCa diagnosis-free survival ($P < 0.001$). **(b)** The cs-PCa diagnosis-free survival curves of the PI-RADS 3 cohort, stratified by initial magnetic resonance imaging-defined prostate-specific antigen density (mPSAD) range. A higher probability of survival was observed in the low mPSAD (<0.15 ng/mL^2) subgroup ($P < 0.001$). **(c)** The cs-PCa diagnosis-free survival curves of the PI-RADS 3 cohort, stratified by prior biopsy status. Biopsy-naïve cases exhibited a lower probability of cs-PCa diagnosis-free survival ($P = 0.018$). **(d)** The cs-PCa diagnosis-free survival curves of the PI-RADS 4–5 cohort, stratified by initial mPSAD range. The subgroup with a high mPSAD (≥0.15 ng/mL^2) demonstrated a lower probability of cs-PCa diagnosis-free survival ($P < 0.001$). **(e)** The cs-PCa diagnosis-free survival curves of the PI-RADS 4–5 cohort, stratified by prior biopsy status. Cases with a history of prior negative biopsy displayed a higher probability of survival without a cs-PCa diagnosis ($P < 0.001$). **(f)** The cs-PCa diagnosis-free survival curves of the PI-RADS 4–5 cohort, stratified by age group. A negative correlation was observed between the age range and the probability of a cs-PCa diagnosis-free survival ($P < 0.001$).

durations of the remaining event-free 1,107 cases, stratified by the initial PI-RADS categories of 1–2, 3, and 4–5, were as follows (format, mean ± standard error): 28.21 ± 1.01, 24.78 ± 0.95, and 16.03 ± 1.25 months, respectively. Notably, among these censored cases, a negative correlation between the initial PI-RADS category and event-free follow-up duration was identified ($P < 0.001$).

The multivariable Cox regression analysis revealed that the initial PI-RADS category, mPSAD, age (cut-off, 60 years), and prior biopsy status significantly affected cs-PCa diagnosis-free survival (Table 2). Figure 3a shows the survival curves of each PI-RADS cohort, demonstrating statistically significant differences among all groups ($P < 0.001$). The cs-PCa diagnosis-free survival durations of PI-RADS 1–2, 3, 4, and 5 cohorts were as follows (format, mean ± standard error): 81.57 ± 1.38, 74.30 ± 1.21, 47.02 ± 2.72, and 21.17 ± 3.41 months, respectively (Table 3).

In the survival analyses, 1, 3, and 5-year cs-PCa diagnosis-free survival rates for the PI-RADS 1–2 cohort were 99.1%, 96.5%, and 93.8%, respectively. Only the initial mPSAD was found to affect cs-PCa diagnosis-free survival in the PI-RADS 1–2 cohort (Table 3).

Table 2. Cox regression of the risk factors for cs-PCa diagnosis in all follow-up cases

Risk factors	Hazard ratio	P value	95% CI
Initial PI-RADS 5*	29.52	<0.001	15.93–54.72
Initial PI-RADS 4*	14.46	<0.001	7.92–26.41
Initial PI-RADS 3*	2.03	0.039	1.04–3.96
Age (>70)**	1.98	0.001	1.37–2.85
Age (60–70)**	1.54	0.013	1.10–2.16
mPSAD ≥0.15 ng/mL²	3.12	<0.001	2.36–4.11
Biopsy-naïve status	1.79	<0.001	1.30–2.46

*Cases with an initial PI-RADS 1–2 category were used as a reference; **cases below the age of 60 years were used as a reference. PI-RADS, Prostate Imaging Reporting and Data System, CI, confidence interval; mPSAD, magnetic resonance imaging-defined prostate-specific antigen density; cs-PCa, clinically significant prostate cancer.

For the PI-RADS 3 cohort, 1, 3, and 5-year cs-PCa diagnosis-free survival rates were 94.9%, 90.9%, and 89.1%, respectively (Table 3). Initial mPSAD and prior biopsy status were the two factors affecting cs-PCa diagnosis-free survival in the PI-RADS 3 cohort (Figure 3b, c). According to Cox multivariable regression analysis, a high mPSAD and biopsy-naïve status were significantly associated with the development of cs-PCa during the follow-up of the PI-RADS 3 cohort [hazard ratio (95% confidence interval): 3.97 (1.92–8.20) and 3.61 (1.35–9.70), respectively].

For the PI-RADS 4 cohort, the 1, 3, and 5-year cs-PCa diagnosis-free survival rates were 56.6%, 55.1%, and 55.1%, respectively. These rates were found to all be 24.2% in the PI-RADS 5 cohort (Table 3). Factor-based evaluation in the combined PI-RADS 4–5 cohort revealed that initial mPSAD, prior biopsy status, and the age group affected cs-PCa diagnosis-free survival (Figure 3d-f).

Considering the 394 cases diagnosed with PCa during follow-up, distributions of ISUP grade groups and NCCN risk groups according to the initial PI-RADS categories are

provided in Table 4. Figure 4 demonstrates the percentages of favorable and unfavorable PCa cases and subgroup analysis results for PI-RADS 3 cases. In PI-RADS 1–2, 3, 4, and 5 cases that were eventually diagnosed with PCa, the initial mp-MRI-to-PCa diagnosis time intervals were observed as median (minimum–maximum) values of 300 (7–2,398) days, 75 (3–1,616) days, 28 (6–1,360) days, and 23 (1–233) days, respectively. Correspondingly, in mean ± standard error format, these intervals were 545.8 ± 121.4, 236.3 ± 38.0, 67.4 ± 12.6, and 32.7 ± 3.2 days, in the given order. Thus, a negative correlation between the PI-RADS category and the time to PCa diagnosis was evident.

Table 3. The cs-PCa diagnosis-free survival results of PI-RADS cohorts and variable-based subgroups

Cohorts and subgroups		Cs-PCa diagnosis-free survival					P values (Log-rank test)
		1 year (%)	3 years (%)	5 years (%)	Survival time (months)		
					Mean ± SE	Median ± SE	
PI-RADS 1–2 cohort		99.1	96.5	93.8	81.57 ± 1.38	-	-
Age (years)	<60	97.8	96.6	96.6	78.55 ± 1.20	-	0.894
	60–70	-	95.7	93.6	78.60 ± 1.39	-	
	>70	-	-	83.3	74.72 ± 5.98	74.10 ± 18.96	
mPSAD (ng/mL²)	<0.15	-	98.7	97.3	84.45 ± 0.78	-	<0.001
	≥0.15	95.2	84.8	70.7	64.68 ± 5.16	74.10 ± -	
Prior negative biopsy	Absent	99.6	96.8	94.8	81.11 ± 2.10	-	0.716
	Present	98.2	95.8	92.2	77.94 ± 1.82	-	
PI-RADS 3 cohort		94.9	90.9	89.1	74.30 ± 1.21	-	-
Age (years)	<60	95.0	95.0	95.0	70.41 ± 1.22	-	0.222
	60–70	96.3	88.8	85.4	73.04 ± 1.94	-	
	>70	88.9	86.0	86.0	60.30 ± 3.25	-	
mPSAD (ng/mL²)	<0.15	97.5	94.1	91.8	75.04 ± 1.18	-	<0.001
	≥0.15	88.4	82.6	82.6	68.43 ± 2.82	-	
Prior negative biopsy	Absent	92.6	88.9	85.4	72.25 ± 1.78	-	0.018
	Present	99.2	94.6	94.6	75.30 ± 1.35	-	
PI-RADS 4 cohort		56.6	55.1	55.1	47.02 ± 2.72	-	-
Age (years)	<60	72.0	72.0	72.0	44.74 ± 3.26	-	0.001
	60–70	56.2	54.6	54.6	46.60 ± 3.90	-	
	>70	38.2	35.8	35.8	26.64 ± 4.69	2.43 ± 1.56	
mPSAD (ng/mL²)	<0.15	67.7	66.4	66.4	56.37 ± 3.27	-	<0.001
	≥0.15	39.1	37.4	-	19.96 ± 2.58	2.10 ± 0.92	
Prior negative biopsy	Absent	52.9	50.6	50.6	41.12 ± 3.12	-	0.141
	Present	65.0	65.0	65.0	54.86 ± 2.72	-	
PI-RADS 5 cohort		24.2	24.2	24.2	21.17 ± 3.41	1.10 ± 0.18	-

PI-RADS, Prostate Imaging Reporting and Data System; SE, standard error; mPSAD, magnetic resonance imaging-defined prostate-specific antigen density; cs-PCa, clinically significant prostate cancer.

Table 4. Histopathological examination results and the prognostic risk groups of cases diagnosed with PCa during follow-up

Initial PI-RADS categories	ISUP grade groups		National Comprehensive Cancer Network risk groups					
	ISUP = 1 PCa (ci-PCa)	ISUP ≥2 PCa (cs-PCa)	Very low	Low	Intermediate (favorable)	Intermediate (unfavorable)	High	Very high
PI-RADS 1-2 (n = 28) (%)	16 (57.1)	12 (42.9)	7 (25.0)	3 (10.7)	5 (17.9)	6 (21.4)	7 (25.0)	0 (0.0)
PI-RADS 3 (n = 80) (%)	50 (62.5)	30 (37.5)	16 (20.0)	17 (21.3)	15 (18.7)	16 (20.0)	14 (17.5)	2 (2.5)
PI-RADS 4, (n = 165) (%)	55 (33.3)	110 (66.7)	11 (6.7)	25 (15.2)	33 (20.0)	48 (29.1)	34 (20.6)	14 (8.5)
PI-RADS 5 (n = 121) (%)	21 (17.4)	100 (82.6)	2 (1.6)	7 (5.8)	5 (4.1)	18 (14.9)	43 (35.6)	46 (38.0)

PI-RADS, Prostate Imaging Reporting and Data System; ISUP, International Society of Urological Pathology; PCa, prostate cancer; ci-PCa: clinically insignificant prostate cancer; cs-PCa, clinically significant prostate cancer; SE, standard error; min, minimum; max, maximum.
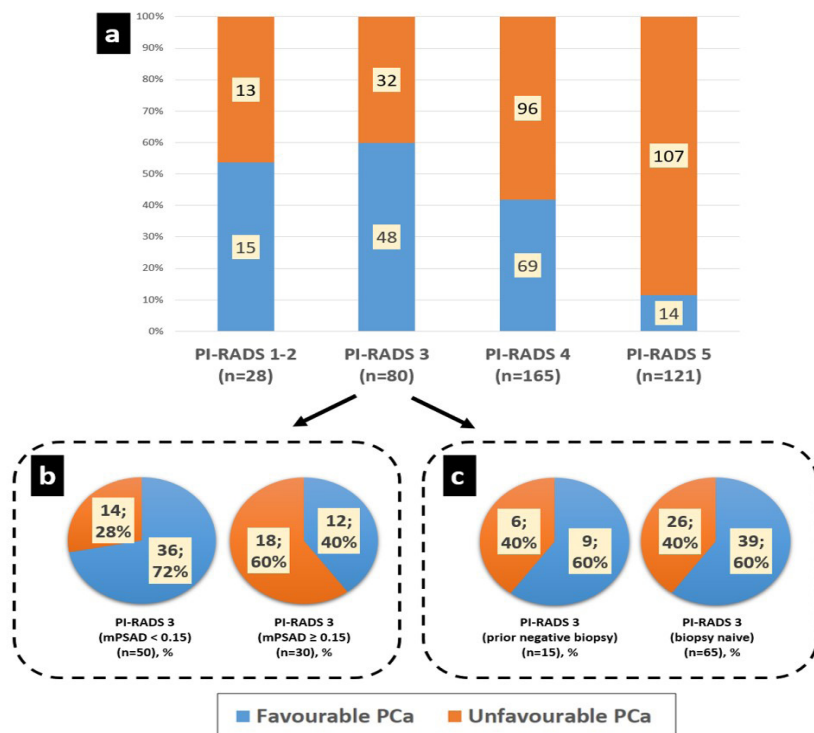
Figure 4. (a) Unfavorable prostate cancer (PCa) percentages in Prostate Imaging Reporting and Data System (PI-RADS) 1–2, 3, 4, and 5 groups were 46%, 40%, 58%, and 88%, respectively. The PI-RADS 1–2 and 3 groups exhibited similar prognostic distribution regarding unfavorable PCa ($P = 0.55$). However, comparing PI-RADS ≤3, 4, and 5 subgroups revealed statistically significant differences concerning unfavorable PCa rates, which were positively correlated with the initial PI-RADS category ($P < 0.001$). **(b, c)** In subgroup analyses, the PI-RADS 3 group was evaluated in terms of magnetic resonance imaging-defined prostate-specific antigen density (mPSAD) range **(b)** and prior biopsy status **(c)** for prognostic assessment. The rate of unfavorable PCa was 28% in the low mPSAD (<0.15 ng/mL^2) subgroup and increased to 60% in the high mPSAD (≥0.15 ng/mL^2) subgroup ($P = 0.005$). Conversely, prior biopsy status had no statistically significant impact on prognostic distribution ($P = 1$).

Very low-risk, low-risk, and intermediate risk-favorable PCa cases according to National Comprehensive Cancer Network (NCCN) classification were accepted as favorable PCa. Intermediate risk-unfavorable, high-risk, and very high-risk PCa cases according to NCCN classification were accepted as unfavorable PCa.

Discussion

Our findings showed that the PI-RADS offered important insights for prognostic evaluation and patient management. As the initial PI-RADS category increased, follow-up losses and time to PCa diagnosis decreased, the probabilities of undergoing biopsy and developing cs-PCa during follow-up increased, and the PCa prognosis worsened. Following the initial PI-RADS category, mPSAD was the second significant variable and was strongly associated with long-term follow-up results. The combination of the PI-RADS and mPSAD could, accordingly, improve diagnostic stratification regarding cs-PCa. According to the cs-PCa diagnosis-free survival analysis, conservative management appeared reasonable for cases with initial PI-RADS 1–2 categories, based on reassuring follow-up results. Similarly, a conservative approach may be advisable in PI-RADS 3 cas-

es, particularly if the initial mPSAD was low (<0.15 ng/mL^2) or if there was a history of prior negative biopsy, both of which provided reassuring follow-up outcomes. Histopathological examination still appeared to be the most reliable approach for cases graded as PI-RADS ≥4, a finding that held even when examining different subgroups of PI-RADS ≥4 cases based on variable clinical factors.

In terms of methodology, we evaluated cases with sufficient follow-up data according to a follow-up scheme that utilized clinico-radiological follow-up findings and histopathological examination results. The clinical and radiological criteria in this scheme were determined through a multidisciplinary approach, considering literature-based evidence and our institutional experience. In event-free cases, the maximum duration without reasonable PCa suspicion or biopsy indication was accepted as the end of fol-

low-up to avoid overestimating cs-PCa diagnosis-free survival.

There are several existing mp-MRI-based follow-up studies. In a study conducted by Venderink et al.[9], cases classified as PI-RADS 1–2 on the follow-up mp-MRI were censored as cs-PCa diagnosis-free at the time of mp-MRI. In a prospective follow-up study conducted by Hauth et al.[10] on PI-RADS 3 and 4 cases, a PI-RADS downgrade (from PI-RADS 3 to 2, and PI-RADS 4 to 3), as well as stable PI-RADS 3 category during follow-up, were considered negative indicators for malignancy. We based our methodology on these two studies to establish the censoring points for radiological follow-up in our study. Additionally, in Hauth et al.'s[10] study, cases with a PI-RADS upgrade (from 3 to 4 and 4 to 5), as well as stable PI-RADS 4 cases during follow-up, were reported to harbor at least a 50% possibility of cs-PCa, emphasizing the necessity for performing a biopsy. Therefore, we decided to consider these cases "indeterminate" unless a biopsy was performed during subsequent follow-up. Conversely, we accepted cases showing a PI-RADS upgrade from PI-RADS 1–2 to 3 as cs-PCa diagnosis-free if follow-up radiological stability was present or at least 1 year of subsequent clinical stability was observed. We based this approach on a prospective study by van der Sar et al.[11], which compared "immediate biopsy" and "close surveillance" approaches for radiologically indeterminate cases, and a literature review by Rivas et al.[12] on the conservative management of indeterminate lesions.

The integration of clinical follow-up with radiological and histopathological findings is one of the important methodological differences in our study. Unlike the studies mentioned above, determining cs-PCa diagnosis-free survival based solely on mp-MRI or biopsy results does not fully reflect clinical practice, as only some patients undergo mp-MRI or biopsy (unless indicated). A significant portion of patients is only clinically monitored with PSA follow-up and DRE, if necessary. Therefore, it can be assumed that patients are cs-PCa diagnosis-free as long as there is clinical stability, no significant increase in follow-up PSA values, no suspicious DRE findings, and no indication for performing a biopsy or mp-MRI. The necessity of this approach was also emphasized in Venderink et al.'s[9] study on cs-PCa diagnosis-free survival of PI-RADS 1–2 cases, suggesting that future studies should focus on longer follow-up periods using a systematic design that includes PSA monitoring and follow-up mp-MRI. However, this integration of clini-

cal follow-up also includes some inevitable uncertainties. The variability of PSA values during follow-up, the subjectivity of DRE, and the lack of standardization for PSA monitoring make this integration methodologically challenging. Nevertheless, different opinions have been reported in the literature about at what point the suspicion of malignancy arises and which PSA kinetics should be used in PSA follow-up, reflecting differences in practice between clinical centers. When establishing the criteria for clinical follow-up stability in our study, we considered the urology literature on PSAV, active surveillance recommendations, and our own clinical experience.

In the 1990s, Carter et al.[8] proposed a PSAV cut-off value of 0.75 ng/mL/year to distinguish cases with and without PCa. It was stated that at least three consecutive PSA values covering a long-term follow-up period were necessary for accurate measurement.[8] Conversely, Venderink et al.[9] reported that cases with a PSA increase of more than 25% during follow-up were referred for further evaluation (mp-MRI or biopsy) in their institution. Regarding active surveillance recommendations, Hefermehl et al.[13] and Hagmann et al.[14] accepted a threshold value of 0.5 ng/mL/year for repeat biopsy during clinical follow-up. In comparison, Nelson et al.[15] reported an optimal cut-off value of 1.18 ng/mL/year for clinical progression in the non-Hispanic white population under active surveillance. Another source states that cases with a PSA doubling time of fewer than 36 months require further evaluation with mp-MRI or biopsy during active surveillance.[16] We considered patients with PSA levels within the age-based normal range or with a PSAV below 0.75 ng/mL/year as clinically stable, as long as there were no suspicious DRE findings. Considering that the majority of our cases had PSA values greater than 3 ng/mL, it is mathematically evident that our PSA monitoring criteria were more stringent compared to the 25% PSA increase criterion mentioned by Venderink et al.[9] and the above-mentioned PSA doubling time criterion recommended for active surveillance.[16] Therefore, the cut-off value we determined appears safer than the aforementioned approaches, except for Hefermehl et al.[13] and Hagmann et al.[14] thresholds. In clinical follow-up evaluation, we sought the presence of at least a 1-year follow-up and at least three PSA measurements to overcome PSA variations that could occur within short time intervals.[8] If the censoring was based solely on clinical monitoring, the follow-up was ended in the presence of procedures such as transurethral resection of the prostate or open prostatectomy, which can significantly lower the PSA level and make it challenging to evaluate clinical stability. Furthermore, in our study, for clinical follow-up to be considered stable in the absence of biopsy or mp-MRI, the final PSA value was required to not exceed 20 ng/mL. We based this criterion on a study conducted by Agnihotri et al.[17], which indicated a high probability of malignancy (above 60%) in cases with PSA values greater than 20 ng/mL, even in the absence of suspicious DRE findings.

The additive impact of mPSAD and PI-RADS category on cs-PCa prediction has been emphasized in several studies.[18,19] Frisbie et al.[20] underlined that these two variables complemented one another in stratifying the risk of cs-PCa. Wang et al.[21], in their study evaluating risk factors associated with progression in patients undergoing active surveillance for PCa, found that both PI-RADS category and PSAD were significant factors in both univariable and multivariable analyses. Ma et al.[22] included age as a variable in their predictive model for cs-PCa, in addition to the PI-RADS category and PSAD, and achieved an AUC of 0.914 after external validation. Patel et al.[23] emphasized the importance of prior biopsy status and reported that the percentages of PCa and cs-PCa in biopsy-naïve cases were approximately twice as high as those in cases with prior negative biopsies. In the same study, after multivariable analyses, age, PSA, PSAD, prostate volume, and PI-RADS 4–5 categories were also found to be significantly associated with cs-PCa. Unlike our study, this study found no statistically significant increase in cs-PCa risk for the PI-RADS 3 group (compared to PI-RADS 1–2 as a reference).[23]

High cs-PCa diagnosis-free survival in our PI-RADS 1–2 cohort was similar to the results reported in the studies by Panebianco et al.[24] and Venderink et al.[9], supporting the low likelihood of detecting cs-PCa during follow-up. In their survival analysis on PI-RADS 1–2 cases, Panebianco et al.[24] found that a 4-year cs-PCa diagnosis-free survival probability was 95% in the biopsy-naïve group and 96% in patients with prior negative biopsies. In Panebianco et al.'s[24] work, PSA was also found to be associated with cs-PCa, in addition to PSAD. Venderink et al.[9] performed survival analysis on 361 PI-RADS 1–2 cases who had a subsequent histopathological examination or follow-up mp-MRI. According to the study results, the 3- and 6-year cs-PCa diagnosis-free survival probabilities in patients with an initial mp-MRI result of PI-RADS 1-2 were 99.6% and 94.1%, respectively.[9] In contrast to our findings, Venderink's et al.[9] study found a significant association between patients' age and the likelihood of cs-PCa diagnosis, while no significant association was found between PSA level, PSAD, or prior biopsy status and the probability of cs-PCa.

Our findings indicate that the probability of cs-PCa development during long-term follow-up in the PI-RADS 3 cohort was approximately 10%. Therefore, the conservative follow-up approach recommended in the literature, comprising PSA monitoring and/or follow-up mp-MRI, may be a reasonable option for managing PI-RADS 3 cases. van der Sar et al.[11] proposed a surveillance strategy for radiologically indeterminate cases, comprising PSA monitoring, if necessary, followed by mp-MRI and, if necessary, delayed biopsy. The majority of patients (57%) preferred this approach over immediate biopsy. No difference in PCa risk profiles was observed between the two approaches. Moreover, a significant portion of PI-RADS 3 cases (39%), where patients selected the conservative pathway, were followed clinically with PSA monitoring only, without the need for follow-up mp-MRI or biopsy, thereby avoiding the risks and costs associated with unnecessary procedures. Rivas et al.[12] similarly stated that surveillance without biopsy may be a viable alternative approach for radiologically indeterminate lesions. Hauth et al.[10], in their prospective study, reported that only 4% of patients with PI-RADS 3 lesions on initial mp-MRI developed cs-PCa during follow-up. In Hauth's et al.[10] work, a follow-up mp-MRI one year later was recommended for PI-RADS 3 cases to exclude the possibility of high-grade cancer development, and the majority of PI-RADS 3 lesions remained stable or decreased in size during follow-up imaging. Similarly, Steinkohl et al.[25] suggested that the ideal timing for follow-up mp-MRI in PI-RADS 3 cases was approximately 12.4 months after the initial examination. Another study on the follow-up of PI-RADS 3 cases indicated that mp-MRI performed 12–24 months later could eliminate the need for biopsy. The same study emphasized the low percentage of cs-PCa in the PI-RADS 3 group (4%) and stated that a PI-RADS upgrade was observed on the pre-biopsy follow-up mp-MRIs in cases eventually diagnosed with cs-PCa.[26] For the PI-RADS 3 cohort, we found that a low mPSAD and the presence of a prior negative biopsy were two protective factors against the development of cs-PCa, based on both univariable and multivariable analyses.

These findings demonstrate that conservative follow-up can be even more reliable in these specific subgroups of PI-RADS 3 cases.

According to 1-year follow-up results, cs-PCa detection in approximately 43% of the PI-RADS 4 cohort and 76% of the PI-RADS 5 cohort reflected the positive predictive value of PI-RADS category aligning with reported cs-PCa detection rates in the literature [59% (39%–78%) and 85% (73%–94%) for PI-RADS 4 and 5 cases, respectively] and emphasized the necessity of a biopsy.[27] The literature recommends an immediate biopsy for PI-RADS 4–5 cases, and even if the initial biopsy is negative, the need for re-biopsy or follow-up mp-MRI is underscored.[10] Meng et al.[28] discussed the follow-up of patients initially categorized as PI-RADS 4–5 with a subsequent nonmalignant targeted biopsy. In follow-up mp-MRIs, a PI-RADS category downgrade was observed in 73% of cases, while PI-RADS 4–5 lesions persisted in 27%. Among cases that were downgraded to PI-RADS 2–3, malignancy was observed in 23%, whereas 62.5% of cases with persistent lesions were diagnosed with cancer.[28] Similarly, Barletta et al. emphasized the high positive predictive value of follow-up mp-MRI and its strong association with the presence of cs-PCa.[29] These studies showed that radiological follow-up could be effective in the diagnostic management of PI-RADS 4–5 cases.

Due to the relatively high percentage of cs-PCa expected in PI-RADS 4–5 cases, there is a paucity of evidence and no widely accepted recommendations regarding conservative follow-up without biopsy. However, it is well-known that not every PI-RADS 4–5 case is malignant, and mimickers such as prostatitis can cause diagnostic confusion.[30] Long-term follow-up findings, such as a PI-RADS category downgrade and PSA regression, may help distinguish between PCa and prostatitis.[31] Therefore, we included 11 patients who had an initial PI-RADS 4–5 mp-MRI but were managed with follow-up mp-MRI and clinical monitoring without undergoing biopsy due to patient preference. In all of these cases, a downgrade to PI-RADS 2–3 was observed on follow-up MRI [median (minimum–maximum) follow-up time, 12.4 (7.00–44.27) months]. Additionally, we included 13 cases with an initial PI-RADS 4 category that were only managed with close clinical monitoring (PSA and DRE), without any biopsy or follow-up mp-MRI, and eventually showed PSA regression [median (minimum-maximum) follow-up time: 36.90 (18.97–50.00) months]. We had no patients in the PI-RADS 5 cohort who were managed only with clinical follow-up. Our study identified advanced age, high mPSAD, and biopsy-naïve status as key risk factors that further increased the likelihood of cs-PCa in the PI-RADS 4–5 cohort. In subgroup analyses of the PI-RADS 4–5 cohort, unlike PI-RADS ≤3 cohorts, cs-PCa diagnosis-free survival probabilities were not reliably high enough to support the conservative follow-up approach.

In our study, PCa cases initially characterized as PI-RADS ≥4 categories were in higher NCCN risk groups, indicating an increased likelihood of definitive treatment requirement and a decreased probability of an active surveillance option compared to PCa cases with initial PI-RADS ≤3. This finding highlights the prognostic value of the initial PI-RADS category. Numerous studies in the literature investigated the relationship between PI-RADS categories and various prognostic factors. Morote et al.[32] demonstrated the association between the PI-RADS group and PCa aggressiveness. Similar to our study, Morote et al.'s[32] work also found that the PI-RADS ≥4 patient group was significantly more associated with aggressive cancers compared to the PI-RADS ≤3 group. Alessi et al.[33] reported that a low PI-RADS score (PI-RADS ≤3) independently excluded the presence of extraprostatic extension with a sensitivity of 99% and a negative predictive value of 98%, irrespective of clinical risk group. Pockros et al.[34] showed that a high PI-RADS category was an independent risk factor for postoperative stage upgrade. The same study reported that lymph node metastasis was only observed in PI-RADS ≥4 cases. In a recent meta-analysis by Rajwa et al.[35], the pre-treatment PI-RADS categories of patients who received definitive local treatment for PCa were found to be associated with post-treatment biochemical recurrence. Another study indicated that the initial PI-RADS category was associated with distant metastasis in intermediate/high-risk PCa cases treated with primary radiation therapy.[36] Another notable finding in our study is the prognostic effect of mPSAD in the PI-RADS 3 cases. The rate of unfavorable PCa in PI-RADS 3 cases with a high mPSAD (≥0.15 ng/mL^2) was similar to that in PI-RADS 4 cases. This finding supports the approach of making a biopsy decision based on a PSAD threshold value of 0.15 ng/mL^2 in PI-RADS 3 cases.[37-41] Indeed, in our study, a significant portion (72%) of PI-RADS 3 cases with a low mPSAD (<0.15 ng/mL^2) were in the favorable PCa group, from a prognostic perspective.

The main strength of our study is the evaluation of histopathological, radiological, and clinical follow-up data from more than 1,300 cases by a multidisciplinary team in a format comparable to a real-life setting.

The limitations of this study include an inability to establish a uniform follow-up protocol due to the research's retrospective design, a heterogeneous dataset spanning 8 years (including variable image acquisition quality on different MRI scanners, improved mp-MRI evaluation over time by different mp-MRI readers, developing MRI-TRUS fusion biopsy experience, and evolving institutional experience regarding management strategies such as mp-MRI referral and clinical follow-up), and potential selection biases that may have occurred due to high follow-up losses in the PI-RADS ≤3 cohorts. Conversely, the comparison of cases in the main dataset with cases that were eligible for survival analysis suggests similarities in various characteristics between the included and excluded cases, offering relative reassurance for having circumvented selection biases.

Another limitation of this study is that the censoring criteria we defined, such as at least ~1 year of clinical stability, PI-RADS 1–2 category on follow-up mp-MRI, radiologically and/or clinically stable PI-RADS 3 cases during follow-up, PI-RADS downgrade to category 3, and even core biopsy results indicating a non-neoplastic pathology or ISUP = 1 PCa, did not provide 100% reassurance for the absence of cs-PCa.[42] However, since it is not feasible to perform whole-mount histopathological examinations, which are considered the gold standard for PCa diagnosis, for most patients, cases were censored based on current evidence and comprehensive clinical judgment. In this way, we aimed to mirror the fundamental clinical approach and increase the applicability of the findings to daily practice.

In conclusion, PI-RADS category significantly influences patient care and offers vital diagnostic and prognostic insights. The combined use of PI-RADS, particularly with mPSAD and other clinical variables, holds promise for serving as a navigational tool for risk stratification and patient management strategies. In the future, multi-center prospective studies with longer follow-up periods and well-standardized follow-up protocols may be able to shed more light on the role and importance of PI-RADS category in patient management.

References

1. Mohler JL, Antonarakis ES, Armstrong AJ, et al. Prostate cancer, version 2.2019, NCCN clinical practice guidelines in oncology. *J Natl Compr Canc Netw*. 2019;17(5):479-505. [CrossRef]

2. Hernandez DJ, Nielsen ME, Han M, Partin AW. Contemporary evaluation of the D'amico risk classification of prostate cancer. *Urology*. 2007;70(5):931-935. [CrossRef]

3. Gupta RT, Mehta KA, Turkbey B, Verma S. PI-RADS: Past, present, and future. *J Magn Reson Imaging*. 2020;52(1):33-53. [CrossRef]

4. Mottet N, van den Bergh RC, Briers E, et al. EAU-EANM-ESTRO-ESUR-SIOG guidelines on prostate cancer-2020 update. Part 1: screening, diagnosis, and local treatment with curative intent. *Eur Urol*. 2021;79(2):243-262. [CrossRef]

5. Scott R, Misser SK, Cioni D, Neri E. PI-RADS v2.1: what has changed and how to report. *SA J Radiol*. 2021;25(1):2062. [CrossRef]

6. Connolly D, Black A, Murray LJ, Napolitano G, Gavin A, Keane PF. Methods of calculating prostate-specific antigen velocity. *Eur Urol*. 2007;52(4):1044-1050. [CrossRef]

7. Oesterling JE, Jacobsen SJ, Chute CG, et al. Serum prostate-specific antigen in a community-based population of healthy men: establishment of age-specific reference ranges. *JAMA*. 1993;270(7):860-864. [CrossRef]

8. Carter HB, Pearson JD, Metter EJ, et al. Longitudinal evaluation of prostate-specific antigen levels in men with and without prostate disease. *JAMA*. 1992;267(16):2215-2220. [CrossRef]

9. Venderink W, van Luijtelaar A, van der Leest M, et al. Multiparametric magnetic resonance imaging and follow-up to avoid prostate biopsy in 4259 men. *BJU Int*. 2019;124(5):775-784. [CrossRef]

10. Hauth E, Jaeger H, Hohmuth H, Beer M. Follow-up MR imaging of PI-RADS 3 and PI-RADS 4 prostate lesions. *Clin Imaging*. 2017;43:64-68. [CrossRef]

11. van der Sar ECA, Kasivisvanathan V, Brizmohun M, et al. Management of radiologically indeterminate magnetic resonance imaging signals in men at risk of prostate cancer. *Eur Urol Focus*. 2019;5(1):62-68. [CrossRef]

12. Rivas JG, Giganti F, Álvarez-Maestro M, et al. Prostate indeterminate lesions on magnetic resonance imaging-biopsy versus surveillance: a literature review. *Eur Urol Focus*. 2019;5(5):799-806. [CrossRef]

13. Hefermehl LJ, Disteldorf D, Lehmann K. Acknowledging unreported problems with active surveillance for prostate cancer: a prospective single-centre observational study. *BMJ Open*. 2016;6(2):e010191. [CrossRef]

14. Hagmann S, Ramakrishnan V, Tamalunas A, et al. Two decades of active surveillance for prostate cancer in a single-center cohort: favorable outcomes after transurethral resection of the prostate. *Cancers*. 2022;14(2):368. [CrossRef]

15. Nelson TJ, Javier-DesLoges J, Deka R, et al. Association of prostate-specific antigen velocity with clinical progression among African American and non-Hispanic white men treated for low-risk prostate cancer with active surveillance. *JAMA Netw Open*. 2021;4(5):e219452. [CrossRef]

16. Punnen S, Carroll MPR, Washington SL. Active surveillance for males with clinically localized prostate cancer. In: Lee WR, Richie JP, Yushak M, eds. UpToDate; 2022. [CrossRef]

17. Agnihotri S, Mittal RD, Kapoor R, Mandhani A. Raising cut-off value of prostate specific antigen (PSA) for biopsy in symptomatic men in India to reduce unnecessary biopsy. *Indian J Med Res*. 2014;139(6):851-856. [CrossRef]

18. Jordan EJ, Fiske C, Zagoria RJ, Westphalen AC. Evaluating the performance of PI-RADS v2 in the non-academic setting. *Abdom Radiol (NY)*. 2017;42:2725-2731. [CrossRef]

19. Wei X, Xu J, Zhong S, et al. Diagnostic value of combining PI-RADS v2.1 with PSAD in clinically significant prostate cancer. *Abdom Radiol (NY)*. 2022;47(10):3574-3582. [CrossRef]

20. Frisbie JW, Van Besien AJ, Lee A, et al. PSA density is complementary to prostate MP-MRI PI-RADS scoring system for risk stratification of clinically significant prostate cancer. *Prostate Cancer Prostatic Dis*. 2023;26(2):347-352. [CrossRef]

21. Wang AZ, O'Connor LP, Yerram N, et al. Association of PI-RADS categories and PSA density with active surveillance progression in patients with prostate cancer. *J Clin Oncol*. 2020;38(6 Suppl):293. [CrossRef]

22. Ma Z, Wang X, Zhang W, et al. Developing a predictive model for clinically significant prostate cancer by combining age, PSA density, and mpMRI. *World J Surg Oncol*. 2023;21(1):83. [CrossRef]

23. Patel HD, Koehne EL, Shea SM, et al. Risk of prostate cancer for men with prior negative biopsies undergoing magnetic resonance imaging compared with biopsy-naive men: a prospective evaluation of the PLUM cohort. *Cancer*. 2022;128(1):75-84. [CrossRef]

24. Panebianco V, Barchetti G, Simone G, et al. Negative multiparametric magnetic resonance imaging for prostate cancer: what's next? *Eur Urol*. 2018;74(1):48-54. [CrossRef]

25. Steinkohl F, Gruber L, Bektic J, et al. Retrospective analysis of the development of PIRADS 3 lesions over time: when is a follow-up MRI reasonable? *World J Urol*. 2018;36(3):367-373. [CrossRef]

26. Boschheidgen M, Schimmöller L, Doerfler S, et al. Single center analysis of an advisable control interval for follow-up of patients with PI-RADS category 3 in multiparametric MRI of the prostate. *Sci Rep*. 2022;12(1):6746. [CrossRef]

27. Oerther B, Engel H, Bamberg F, Sigle A, Gratzke C, Benndorf M. Cancer detection rates of the PI-RADSv2.1 assessment categories: systematic review and meta-analysis on lesion level and patient level. *Prostate Cancer Prostatic Dis*. 2022;25(2):256-263. [CrossRef]

28. Meng X, Chao B, Chen F, Huang R, Taneja SS, Deng FM. Followup of men with PI-RADS™ 4 or 5 abnormality on prostate magnetic resonance imaging and nonmalignant pathological findings on initial targeted prostate biopsy. *J Urol*. 2021;205(3):748-754. [CrossRef]

29. Barletta F, Stabile A, Mazzone E, et al. How to optimize follow-up in patients with a suspicious multiparametric MRI and a subsequent negative targeted prostate biopsy. Results from a large, single-institution series. *Urol Oncol*. 2022;40(3):103. [CrossRef]

30. Uysal A, Karaosmanoğlu AD, Karcaaltıncaba M, et al. Prostatitis, the great mimicker of prostate cancer: can we differentiate them quantitatively with multiparametric MRI? *AJR Am J Roentgenol*. 2020;215(5):1104-1112. [CrossRef]

31. Azab S, Osama A, Rafaat M. Does normalizing PSA after successful treatment of chronic prostatitis with high PSA value exclude prostatic biopsy? *Transl Androl Urol*. 2012;1(3):148-152. [CrossRef]

32. Morote J, Borque-Fernando A, Triquell M, et al. Multiparametric magnetic resonance imaging grades the aggressiveness of prostate cancer. *Cancers (Basel)*. 2022;14(7):1828. [CrossRef]

33. Alessi S, Pricolo P, Summers P, et al. Low PI-RADS assessment category excludes extraprostatic extension (≥ pT3a) of

prostate cancer: a histology-validated study including 301 operated patients. *Eur Radiol*. 2019;29(10):5478-5487. [CrossRef]

34. Pockros B, Stensland KD, Parries M, Frankenberger E, Canes D, Moinzadeh A. Preoperative MRI PI-RADS scores are associated with prostate cancer upstaging on surgical pathology. *Prostate*. 2022;82(3):352-358. [CrossRef]

35. Rajwa P, Mori K, Huebner NA, et al. The prognostic association of prostate MRI PI-RADS™ v2 assessment category and risk of biochemical recurrence after definitive local therapy for prostate cancer: a systematic review and meta-analysis. *J Urol*. 2021;206(3):507-516. [CrossRef]

36. Turchan WT, Kauffmann G, Patel P, Oto A, Liauw SL. PI-RADS score is associated with biochemical control and distant metastasis in men with intermediate-risk and high-risk prostate cancer treated with radiation therapy. *Urol Oncol*. 2020;38(6):600. [CrossRef]

37. Venderink W, van Luijtelaar A, Bomers JGR, et al. Results of targeted biopsy in men with magnetic resonance imaging lesions classified equivocal, likely or highly likely to be clinically significant prostate cancer. *Eur Urol*. 2018;73(3):353-360. [CrossRef]

38. Messina E, Pecoraro M, Laschena L, et al. Low cancer yield in PI-RADS 3 upgraded to 4 by dynamic contrast-enhanced MRI: is it time to reconsider scoring categorization? *Eur Radiol*. 2023;33(8):5828-5839. [CrossRef]

39. Vourganti S, Rastinehad A, Yerram NK, et al. Multiparametric magnetic resonance imaging and ultrasound fusion biopsy detect prostate cancer in patients with prior negative transrectal ultrasound biopsies. *J Urol*. 2012;188(6):2152-2157. [CrossRef]

40. Distler FA, Radtke JP, Bonekamp D, et al. The value of PSA density in combination with PI-RADS™ for the accuracy of prostate cancer prediction. *J Urol*. 2017;198(3):575-582. [CrossRef]

41. Washino S, Okochi T, Saito K, et al. Combination of prostate imaging reporting and data system (PI-RADS) score and prostate-specific antigen (PSA) density predicts biopsy outcome in prostate biopsy naïve patients. *BJU Int*. 2017;119(2):225-233. [CrossRef]

42. Ahmed HU, Bosaily AES, Brown LC, et al. Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study. *Lancet*. 2017;389(10071):815-822. [CrossRef]

Supplementary Table 1. Multiparametric prostate MRI protocols of the two most commonly used devices at our center

1.5 T Philips Achieva dStream multiparametric prostate MRI protocol						
Sequences and parameters	T2W axial	T2W coronal	T2W sagittal	DWI (large FOV)	DWI (focus)	DCE
FOV (mm x mm)	140 x 140	140 x 140	250 x 250	250 x 215	150 x 150	220 x 303
Matrix	232 x 217	232 x 208	208 x 208	84 x 69	64 x 61	128 x 178
Slice thickness (mm)	3	3	4	5	4	3
TR (ms)	6000	2900	3080	4350	5000	4.9
TE (ms)	100	100	120	85	98	2.3
Flip angle (^{0})	90	90	90	90	90	8
b value (s/mm^2)	-	-	-	0 and 800	0, 800, and 1500	-
Average	2	2	2	6	4	1
Number of phases	-	-	-	-	-	27
3.0 T GE SIGNA™ architect multiparametric prostate MRI protocol						
Sequences and parameters	T2W axial	T2W coronal	T2W sagittal	DWI (large FOV)	DWI (focus)	DCE
FOV (mm x mm)	200 x 200	180 x 180	180 x 180	340 x 340	200 x 200	240 x 240
Matrix	452 x 318	452 x 288	452 x 288	120 x 134	92 x 46	224 x 200
Slice thickness (mm)	3	3	3	5	3	3
TR (ms)	3900	3000	3180	8500	4800	3.6
TE (ms)	140	136	136	Minimum	Minimum	Minimum
Flip angle (^{0})	165	165	165	-	-	12
b value (s/mm^2)	-	-	-	50 and 800	50, 800, and synthetic 1400	-
Average	1	1.5	1.5	2 and 4	4 and 12	1
Number of phases	-	-	-	-	-	30

During DCE examination, gadolinium-based intravenous contrast agent is administered at a concentration of 0.1-0.2 mmol/kg and with an injection rate of 2–4 mL/s. T2W, T2-weighted, DWI, diffusion-weighted imaging, DCE, dynamic contrast-enhanced, FOV, field of view, TR, repetition time, TE, echo time.
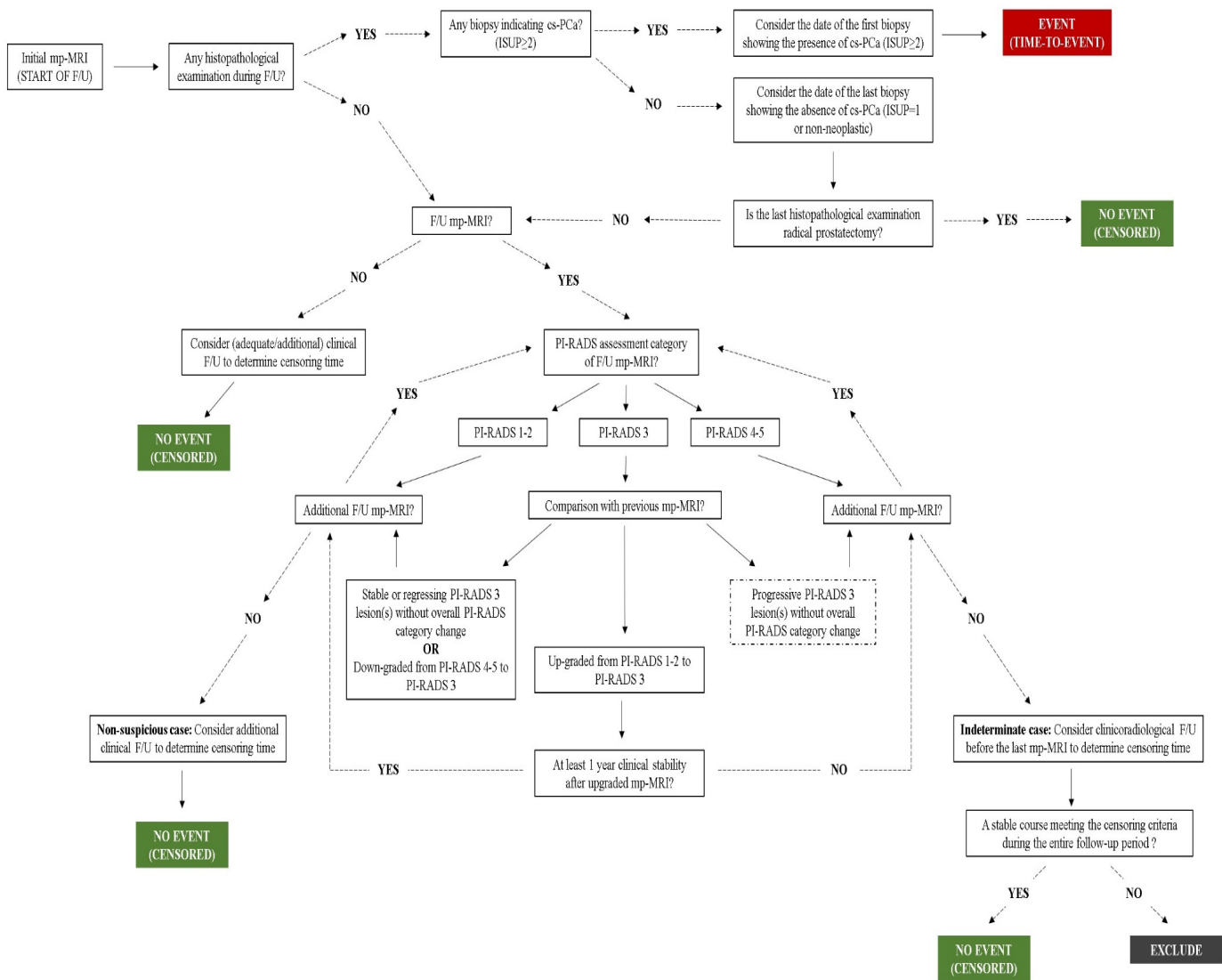
Supplementary Figure 1. Schematic algorithm for follow-up evaluation. ISUP, International Society of Urological Pathology; cs-PCa, clinically significant prostate cancer; mp-MRI, multiparametric prostate magnetic resonance imaging; PI-RADS, Prostate Imaging Reporting and Data System.
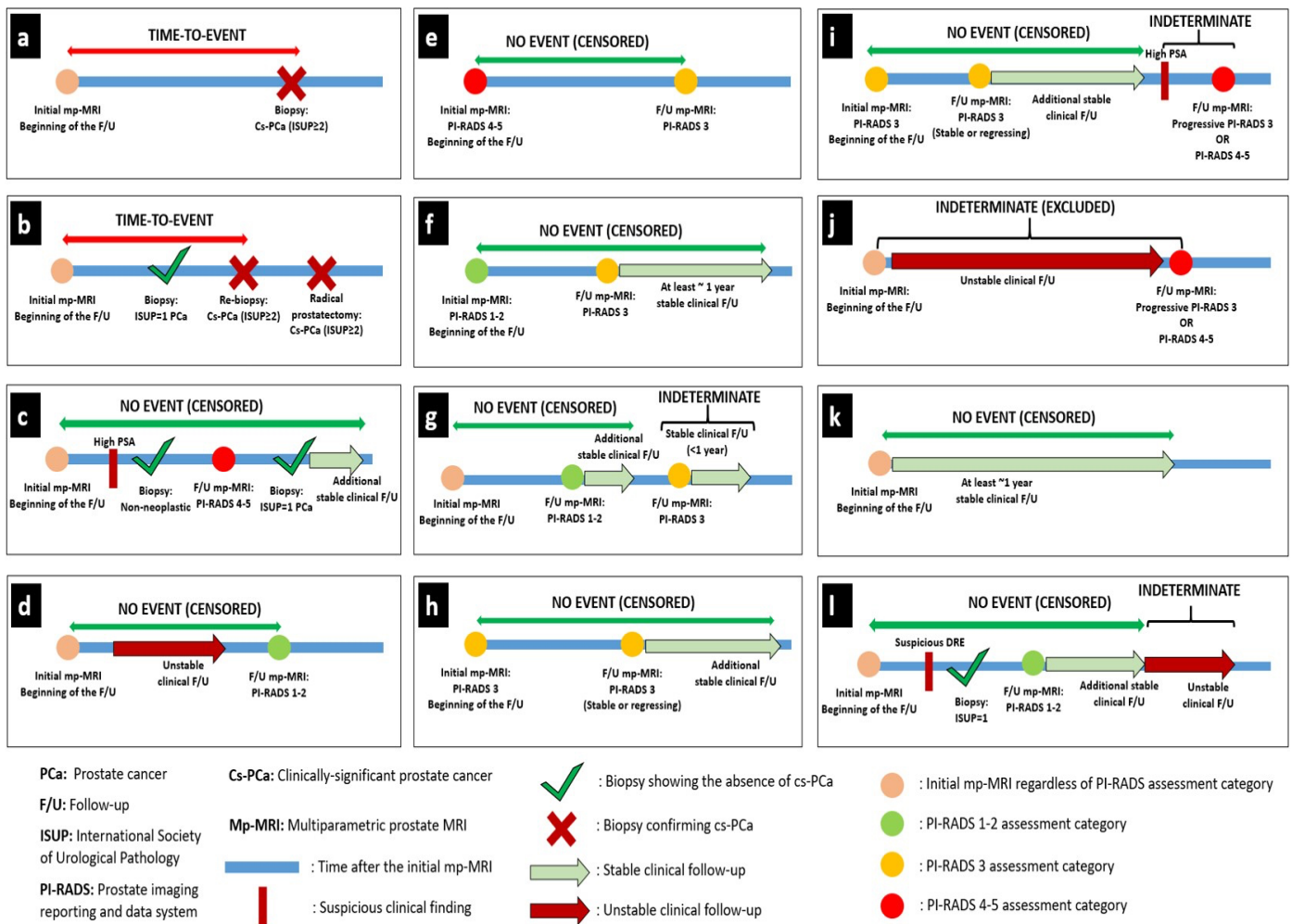
Supplementary Figure 2. Illustrative case examples explaining follow-up evaluation. **(a, b)** The time-to-event was calculated considering the first detection time of cs-PCa (ISUP ≥2). **(c)** Cases without cs-PCa diagnosis during follow-up were considered cs-PCa-diagnosis-free at the points where histopathological examination results indicated non-neoplastic pathology or ISUP = 1 PCa, regardless of interim clinical or radiological follow-up findings. After biopsy, the final censoring time was determined based on additional clinical and/or radiological stability, if available. In the absence of any histopathological examination during an evaluated follow-up interval or in the subsequent follow-up period after a histopathological examination confirming the absence of cs-PCa, radiological stability was examined first and then clinical stability was considered to determine the final censoring time. **(d)** Cases with follow-up PI-RADS category 1–2 were considered cs-PCa diagnosis-free at the time of the follow-up mp-MRI, independent of previous clinical follow-up findings. After mp-MRI, the final censoring time was determined based on additional clinical and radiological follow-up, if available.

(e-i) The approach to cases with follow-up PI-RADS category 3 was determined based on the previous PI-RADS. **(e)** Cases with a PI-RADS downgrade from PI-RADS 4–5 to PI-RADS 3 were considered cs-PCa diagnosis-free at the time of follow-up mp-MRI. The final censoring time was determined based on additional clinical and radiological follow-up, if available. **(f)** Cases with a PI-RADS upgrade from PI-RADS 1–2 to PI-RADS 3 were evaluated for the presence of at least 1-year stable subsequent clinical follow-up or further radiological stability to be considered as cs-PCa diagnosis-free. **(g)** PI-RADS 3 cases that did not meet these criteria were considered indeterminate at the end of follow-up. Censoring was done based on the stable clinical and/or radiological follow-up interval between the first mp-MRI and the indeterminate follow-up interval. Cases without such a stable follow-up interval were excluded from survival analysis. **(h)** PI-RADS 3 cases that did not show any PI-RADS change and radiological progression in the follow-up mp-MRI were considered as cs-PCa-diagnosis-free at the time of follow-up mp-MRI. The final censoring time was determined based on additional clinical and radiological follow-up, if available. **(i)** Cases with progressive PI-RADS 3 lesion(s) or follow-up PI-RADS category of 4–5 were considered indeterminate regardless of interim clinical follow-up, unless a subsequent histopathological examination was performed. Censoring was done based on the stable clinical and/or radiological follow-up interval between the first mp-MRI and the indeterminate follow-up interval. **(j)** Cases without such a follow-up interval were excluded from survival analysis. **(k)** In the absence of histopathological or radiological examination during the evaluated follow-up interval or in the subsequent follow-up period after these examinations, clinical stability was evaluated to determine the final censoring time. Cases showing at least 1 year of clinical stability, regardless of baseline PI-RADS category, were censored as cs-PCa diagnosis-free. **(l)** Cases where the criteria for clinical stability were not met during a certain time interval were considered indeterminate. Censoring was done based on the stable clinical interval between the first mp-MRI and the indeterminate follow-up interval. cs-PCa, clinically significant prostate cancer; ISUP, International Society of Urological Pathology; PCa, prostate cancer; PI-RADS, Prostate Imaging Reporting and Data System; mp-MRI, multiparametric prostate magnetic resonance imaging

Influence of image preprocessing on the segmentation-based reproducibility of radiomic features: *in vivo* experiments on discretization and resampling parameters

Burak Koçak

Sabahattin Yüzkan

Samet Mutlu

Mehmet Karagülle

Ahmet Kala

Mehmet Kadıoğlu

Sıla Solak

Şeyma Sunman

Zişan Hayriye Temiz

Ali Kürşad Ganiyusufoğlu

University of Health Sciences, Başakşehir Çam and Sakura City Hospital, Clinic of Radiology, İstanbul, Türkiye

PURPOSE

To systematically investigate the impact of image preprocessing parameters on the segmentation-based reproducibility of magnetic resonance imaging (MRI) radiomic features.

METHODS

The MRI scans of 50 patients were included from the multi-institutional Brain Tumor Segmentation 2021 public glioma dataset. Whole tumor volumes were manually segmented by two independent readers, with the participation of eight readers. Radiomic features were extracted from two sequences: T2-weighted (T2) and contrast-enhanced T1-weighted (T1ce). Two methods were considered for discretization: bin count (i.e., relative discretization) and bin width (i.e., absolute discretization). Ten discretization (five for each method) and five resampling parameters were varied while other parameters were fixed. The intraclass correlation coefficient (ICC) was used for reliability analysis based on two commonly used cut-off values (0.75 and 0.90).

RESULTS

Image preprocessing parameters had a significant impact on the segmentation-based reproducibility of radiomic features. The bin width method yielded more reproducible features than the bin count method. In discretization experiments using the bin width on both sequences, according to the ICC cut-off values of 0.75 and 0.90, the rate of reproducible features ranged from 70% to 84% and from 35% to 57%, respectively, with an increasing percentage trend as parameter values decreased (from 84 to 5 for T2; 100 to 6 for T1ce). In the resampling experiments, these ranged from 53% to 74% and from 10% to 20%, respectively, with an increasing percentage trend from lower to higher parameter values (physical voxel size; from 1 x 1 x 1 to 2 x 2 x 2 mm³).

CONCLUSION

The segmentation-based reproducibility of radiomic features appears to be substantially influenced by discretization and resampling parameters. Our findings indicate that the bin width method should be used for discretization and lower bin width and higher resampling values should be used to allow more reproducible features.

KEYWORDS

Reproducibility, preprocessing, radiomics, texture analysis, biomarker

Corresponding author: Burak Koçak

E-mail: drburakkocak@gmail.com

Radiomics is a field of medical image analysis that enables the digital decoding of images into high-throughput quantitative features.[1] Medical images may contain hidden patterns, indicating the underlying pathophysiology of the examined tissue. Based on this assumption, radiomic features derived from these images might help characterize tissues and guide clinical decision-making.[1,2] Support for this notion has arisen from numerous studies that have addressed the capability of radiomics in making predictions regarding different clinical endpoints.[3] There has been an exponential increase in publications related to

radiomics, with a yearly growth rate of 19.6% and a doubling time of 3.9 years.[4] However, reproducing and validating published studies is still challenging due to a lack of standardized definitions, parameter settings, and inadequate reporting.[5-9]

Before implementing radiomics in clinical practice, it is necessary to have a thorough understanding of the reproducibility of radiomic features. Many previous publications have emphasized the dependency of radiomic features on different factors, such as temporal variability,[10,11] scanning parameters,[12-14] delineation uncertainty,[15,16] reconstruction algorithms,[17] preprocessing,[8] and organ motion.[18] The absolute value and statistical distribution of the radiomics features are considerably affected by the aforementioned determinants, which in turn affects the robustness and generalizability of any subsequent analysis derived from these features. To overcome this divergence, the Image Biomarker Standardization Initiative (IBSI) attempted to standardize the radiomic feature extraction process, focusing on the issues of repeatability, reproducibility, and validation in quantitative image analysis and radiomics.[5] According to this initiative, standardized image processing should be performed before radiomic feature extraction.[5] Nonetheless, no specific processing parameter settings have been published to date, which underlines the requirement for additional research.[8,19,20]

One of the most important steps in the radiomic pipeline that affects reproducibility is segmentation or delineation.[21,22] For example, a feature might be highly reproducible in a test–retest setting, but there is no guarantee that this feature will be robust after segmentation. Segmentation-based reproducibility analysis is extensively used to reduce the high dimensionality of radiomics data as a data handling step for subsequent predictive modeling procedures.[2,23] However, only a limited number of studies have focused on the impact of preprocessing settings on segmentation-based feature reproducibility.[24,25] Duron et al.[24] studied magnetic resonance imaging (MRI)-based radiomic features of lachrymal gland tumors and breast lesions with a focus on discretization techniques. Lu et al.[25] investigated positron emission tomography/computed tomography (PET/CT)-based radiomic features in patients with nasopharyngeal carcinoma, again with a focus on discretization. No research has specifically assessed the impact of both image voxel resampling and gray-level discretization on the segmentation-based reproducibility of the radiomic features. However, these two preprocessing methods are frequently encountered in radiomic feature extraction software tools.

The purpose of this study was to systematically investigate the effect of image preprocessing parameters on the segmentation-based reproducibility of radiomic features from MRI and to recommend reasonable parameter settings for achieving highly reproducible features.

Methods

Figure 1 depicts the key study steps to help readers understand the methodology.

Dataset

In this study, we used the Brain Tumor Segmentation (BraTS) 2021 public glioma dataset,[26-28] which does not require local ethical approval. The MRI data for the BraTS 2021 challenge were collected using various clinical protocols and scanners from a variety of data-contributing institutions. There were four MRI sequences in the dataset: T1-weighted (T1), T2-weighted (T2), contrast-enhanced T1-weighted (T1ce), and fluid-attenuated inversion recovery (FLAIR). All BraTS MRI scans underwent standardized preprocessing, which included the conversion of Digital Imaging and Communications in Medicine-format files to Neuroimaging Informatics Technology Initiative format, co-registration to the same anatomical template (SRI24),[29] isotropic voxel resampling (1 x 1 x 1 mm³), and skull-stripping.[30]

For this reproducibility study, 50 patients with gliomas were randomly selected. Patient identifiers are provided in the Supplementary Table S1. Readers who performed the segmentation used all four sequences. Only two sequences-T2 and T1ce-were used for the preprocessing experiments to assess the dependency of the results on the different sequences; the use of more sequences may have become unfeasible considering the workload and complexity of the study. The T2 sequence was selected to represent the outermost boundary of the tumor, and T1ce was used to evaluate the radiomic features on a different image contrast, considering the relatively homogeneous appearance of glial tumors in T2 compared with T1ce.

Main points

- Variations of image preprocessing parameters, regarding discretization and resampling, have a significant impact on the segmentation-based reproducibility of radiomic features.

- The bin width method yields more reproducible features than the bin count method for discretization.

- Using lower bin width values and higher resampling values could help produce more reproducible features.

- The optimal preprocessing parameters should be determined within the radiomic pipeline.

- To allow replication, preprocessing parameters should be transparently reported in radiomic publications due to their importance.
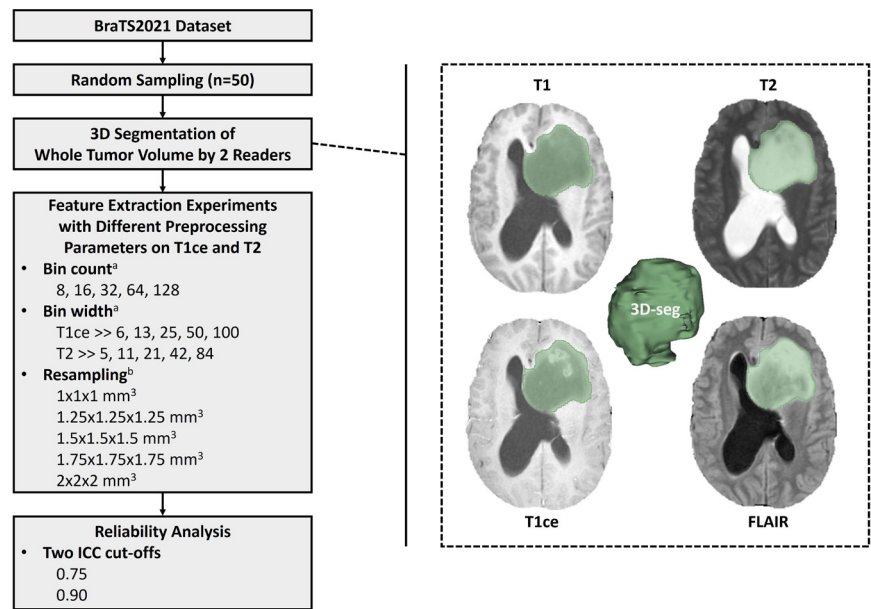
Figure 1. Key study steps and segmentation approach. 3D, three-dimensional; ICC, intraclass correlation coefficient; T1, T1-weighted; T2, T2-weighted; T1ce, contrast-enhanced T1-weighted; FLAIR, fluid-attenuated inversion recovery; 3D-seg, three-dimensional segmentation. [a]Resampling fixed to 1 x 1 x 1 mm³. [b]Discretization fixed to a bin count of 32.

Segmentation

The glial tumors were manually segmented using 3D Slicer software v4.11. The pathological high signal intensity that appears in T2 and FLAIR sequences was used to segment the entire tumor volume. Readers were also free to use any of the four sequences available in the dataset to determine tumor borders (T1, T2, FLAIR, and T1ce). Figure 1 also illustrates the segmentation approach.

The segmentation process involved eight readers (three radiologists and five radiology residents), with two readers (one radiology specialist and one radiology resident) for each patient. All of the specialists worked in the neuroradiology division. Two of these had ≥3 years and one had ≥1 years of experience in neuroimaging as a specialist. During the study, all of the residents were in their second or third year in radiology and on their first neuroradiology rotation.

Preprocessing

All images were normalized to a scale of 100 based on the mean and standard deviation (SD) of voxel intensity values. To avoid negative values, the voxel arrays were shifted by 300.

Experiments were conducted by changing the discretization and resampling parameters. For discretization, two methods were considered: bin count (i.e., relative discretization) and bin width (i.e., absolute discretization). The following preprocessing parameters were used for bin count: 8, 16, 32, 64, and 128. For the bin width method, the following preprocessing settings were used for T1ce: 6, 13, 25, 50, and 100; for T2: 5, 11, 21, 42, and 84. The bin width values were determined based on the first-order range in the dataset to get an approximately equal number of gray levels compared with the bin count approach. When experimenting with the above-mentioned two discretization approaches, the resampling parameter was fixed to 1 x 1 x 1 mm^3. For resampling, the physical voxel sizes were rescaled to 1 x 1 x 1, 1.25 x 1.25 x 1.25, 1.5 x 1.5 x 1.5, 1.75 x 1.75 x 1.75, and 2 x 2 x 2 mm^3. When performing the resampling experiments, the discretization parameter was fixed to a bin count of 32.

Feature extraction

Three-dimensional radiomic features, including shape and texture, were extracted in batch mode using the PyRadiomics open-source software environment (PyRadiomics v3.0.1; NumPy v1.23.5; SimpleITK v2.3.0;

PyWavelet v1.4.1; Python 3.10.12).[31] The total number of features in each sequence was 1.106. Original, Laplacian of Gaussian (LoG)-filtered, and wavelet-transformed images were used in the feature extraction. The LoG filtering was performed with sigma values of 2, 4, and 6 mm, corresponding to fine, medium, and coarse patterns. The main feature classes were shape, first order, gray-level co-occurrence matrix, gray-level size zone matrix, gray-level run-length matrix, gray-level dependence matrix, and neighboring gray-tone difference matrix.

Statistical analysis

The R v4.3 (rstatix v0.7.2) and Python v3.7 (pingouin v0.5.2) software packages were utilized to conduct statistical analyses. To measure feature reproducibility, the intraclass correlation coefficient (ICC) was estimated based on two-way random effects, absolute agreement, and single measurement, under the Shrout and Fleiss convention.[32] The interpretation scale for the ICC was as follows: ICC <0.50, poor; 0.50≤ ICC <0.75, moderate; 0.75≤ ICC <0.90, good; and ICC ≥0.90, excellent.[33] Two thresholds-0.75 and 0.90-were used to report the percentages of reproducible features. The normality of the ICC values was determined using the Shapiro–Wilk test. Depending on the group

distributions, paired tests, notably the one-way repeated measures analysis of variance (ANOVA) and the student t-test, were used to evaluate statistical differences in continuous variables for all and pair-wise comparisons, respectively. McNemar's test was utilized to compare the distribution of categorical variables (i.e., reproducible vs. non-reproducible features based on ICC cut-off values). Statistical results were considered significant if P values were ≤0.05. Multiple comparisons were subjected to multiplicity correction using the Tukey test or Bonferroni correction as appropriate. In these comparisons, statistical significance was determined based on adjusted or unadjusted but corrected P values, for the Tukey test and Bonferroni correction, respectively.

Results

Figure 2 presents the distribution of the ICC estimates for various preprocessing processes, including discretization with bin count, discretization with bin width, and voxel resampling. Detailed descriptive statistics of the ICC estimates based on preprocessing processes are presented in Table 1. As the bin width was reduced in the experiments, the mean ICC values increased. In experiments involving the bin count, the mean ICC val-
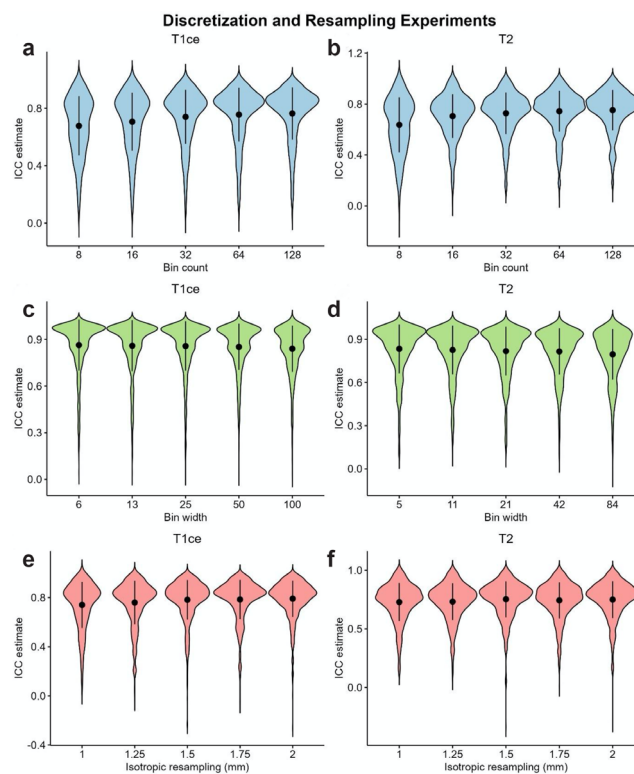
Figure 2. Distribution of the intraclass correlation coefficient (ICC) estimates for different preprocessing steps. Experiments with bin count **(a, b)**, bin width **(c, d)**, and voxel resampling **(e, f)** on contrast-enhanced T1-weighted (T1ce) and T2-weighted (T2) sequences. The filled circle and bar inside the violin represent the mean and standard deviation, respectively.

ues increased as the bin count increased. Both tests revealed that an increase in the number of gray levels led to an increase in the mean ICC values and, in turn, the segmentation-based reproducibility of radiomic features. The mean ICC values were statistically significantly different and higher in the bin width group (for T1ce, mean ± SD, 0.855 ± 0.158; for T2, mean ± SD, 0.818 ± 0.169) than in the bin count group (for T1ce, mean ± SD, 0.729 ± 0.196; for T2, mean ± SD, 0.713 ± 0.180) on both of the T1ce [$t(2,764) = -28.2$, $P < 0.001$] and T2 [$t(2,764) = -22.3$, $P < 0.001$] sequences. For the resampling, the mean ICC values improved as the resampled physical voxel size increased.

Table 2 presents the ANOVA findings for parameter differences across experimental groups. Although the effect sizes were minor (range: 0.002–0.029), all comparisons for all three preprocessing experiments were statistically significant ($P < 0.001$ for all experiments in both sequences). The statistically significant pairs following the post-hoc Tukey test are summarized in Table 3. Considering all evaluations based on sequence and preprocessing experiments, there were statistically significant differences at least between all minimum and maximum numeric values of the preprocessing parameters (e.g., bin count of 8 vs. 128; resampling 1 x 1 x 1 vs. 2 x 2 x 2 mm³).

Figures 3 and 4 depict the percentages of features with good and excellent reproducibility in the discretization and resampling experiments, based on two typical ICC cut-off values (0.75 and 0.90). In the discretization experiments with bin count on both sequences, taking the ICC cut-offs of 0.75 and 0.90 into account, the rate of reproducible features was 36%–69% and 9%–19%, respectively, with an increasing percentage trend from lower parameter values to higher parameter values. In the discretization experiments with bin width on two sequences, with the ICC cut-off values of 0.75 and 0.90, the rate of reproducible features was 70%–84% and 35%–57%, respectively, with an increasing percentage trend as parameter values decreased. In resampling experiments on both sequences, with the ICC cut-off values of 0.75 and 0.90, the rate of reproducible features was 53%–74% and 10%–20%, respectively, with an increasing percentage trend from lower to higher parameter values.

Given a fixed first-order range in a sequence calculated based on the dataset, the bin width experiments outperformed the re-

Table 1. Descriptive statistics of intraclass correlation coefficients for preprocessing experiments

Sequence	Preprocessing method	Preprocessing parameter	ICC estimate	
			Mean	SD
T1ce	Bin count	8	0.678	0.206
		16	0.707	0.203
		32	0.740	0.188
		64	0.757	0.187
		128	0.765	0.182
	Bin width	6	0.864*	0.165
		13	0.859*	0.165
		25	0.858*	0.160
		50	0.852*	0.149
		100	0.840*	0.149
	Resampling	1 x 1 x 1 mm³	0.740	0.188
		1.25 x 1.25 x 1.25 mm³	0.760	0.178
		1.5 x 1.5 x 1.5 mm³	0.782	0.161
		1.75 x 1.75 x 1.75 mm³	0.784	0.160
		2 x 2 x 2 mm³	0.791	0.147
T2	Bin count	8	0.637	0.216
		16	0.705	0.172
		32	0.728	0.163
		64	0.743	0.160
		128	0.752	0.158
	Bin width	5	0.834*	0.170
		11	0.826*	0.169
		21	0.818*	0.170
		42	0.816*	0.160
		84	0.796*	0.175
	Resampling	1 x 1 x 1 mm³	0.728	0.163
		1.25 x 1.25 x 1.25 mm³	0.731	0.157
		1.5 x 1.5 x 1.5 mm³	0.752	0.154
		1.75 x 1.75 x 1.75 mm³	0.743	0.155
		2 x 2 x 2 mm³	0.749	0.157

*Top 10 values. ICC, intraclass correlation coefficient; SD, standard deviation; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted.

Table 2. One-way repeated measures analysis of variance results of intraclass correlation coefficients

Preprocessing method	Sequence	dfN	dfD	F	Generalized eta-squared	P
Bin width	T1ce	1	5528	14.722	0.003	<0.001
	T2	1	5528	29.810	0.005	<0.001
Bin count	T1ce	1	5528	105.387	0.019	<0.001
	T2	1	5528	163.082	0.029	<0.001
Resampling	T1ce	1	5528	62.761	0.011	<0.001
	T2	1	5528	12.780	0.002	<0.001

T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; dfN, degrees of freedom in the numerator; dfD, degrees of freedom in the denominator.

Table 3. Statistically significant pairs after post-hoc Tukey test for one-way repeated measures analysis of variance

Preprocessing method	Sequence	Preprocessing parameters		Estimate	95% CI lower	95% CI upper	Adjusted *P*
		Group#1	Group#2				
Bin width	T1ce	6	100	−0.024	−0.042	−0.006	0.004
		13	100	−0.019	−0.038	−0.001	0.032
	T2	5	84	−0.038	−0.058	−0.018	<0.001
		11	84	−0.030	−0.050	−0.011	<0.001
		21	84	−0.022	−0.042	−0.003	0.016
		42	84	−0.020	−0.040	<0.001	0.044
Bin count	T1ce	8	16	0.029	0.006	0.051	0.005
		8	32	0.062	0.040	0.085	<0.001
		8	64	0.079	0.056	0.101	<0.001
		8	128	0.087	0.064	0.109	<0.001
		16	32	0.034	0.011	0.056	<0.001
		16	64	0.050	0.028	0.072	<0.001
		16	128	0.058	0.035	0.080	<0.001
		32	128	0.024	0.002	0.047	0.028
	T2	8	16	0.068	0.047	0.088	<0.001
		8	32	0.091	0.071	0.111	<0.001
		8	64	0.106	0.086	0.126	<0.001
		8	128	0.115	0.094	0.135	<0.001
		16	32	0.023	0.003	0.044	0.016
		16	64	0.038	0.018	0.059	<0.001
		16	128	0.047	0.027	0.067	<0.001
		32	128	0.024	0.003	0.044	0.013
Resampling[1]	T1ce	1	1.25	0.020	<0.001	0.039	0.047
		1	1.5	0.042	0.022	0.061	<0.001
		1	1.75	0.044	0.025	0.063	<0.001
		1	2	0.051	0.031	0.070	<0.001
		1.25	1.5	0.022	0.003	0.041	0.017
		1.25	1.75	0.024	0.005	0.044	0.006
		1.25	2	0.031	0.012	0.051	<0.001
	T2	1	1.5	0.024	0.006	0.043	0.003
		1	2	0.021	0.003	0.039	0.014
		1.25	1.5	0.021	0.003	0.039	0.016

[1]Performed with isotropic fashion. One dimension (mm) of the voxel is presented. CI, confidence interval; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted.

spective bin count (e.g., for T1ce, a bin count of 128 vs. a bin width of 6) in terms of the percentages of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility in all comparisons, with statistically significant distributional differences (Table 4).

Figures 5 and 6 for the T1ce sequence and Supplementary Figures S1 and S2 for the T2 sequence depict the reproducibility of radiomic features according to the feature classes and image types from which they were extracted. In the qualitative evaluation of these bar charts, there was no major trend deviation other than the original image against the general trend.

Discussion

In this study, we systematically investigated the influence of image preprocessing parameters (i.e., discretization and resampling) on the segmentation-based reproducibility of MRI radiomic features and found a significant impact. The bin width method yielded more reliable features than the bin count method. Using lower bin width values and higher resampling values produced more reproducible features.

Several studies have evaluated the influence of preprocessing and segmentation independently,[34] neglecting their influence on

each other to a large extent. To our knowledge, very few studies have focused on the impact of preprocessing settings on segmentation-based reproducibility.[24,25] Additionally, no research has specifically assessed the impact of both image voxel resampling and gray-level discretization on the segmentation-based reproducibility of radiomic features.

Duron et al.[24] studied two independent MRI datasets of lachrymal gland tumors and breast lesions from two different centers, with two-dimensional delineations for each dataset. They evaluated six absolute (i.e., fixed bin width method) and eight relative
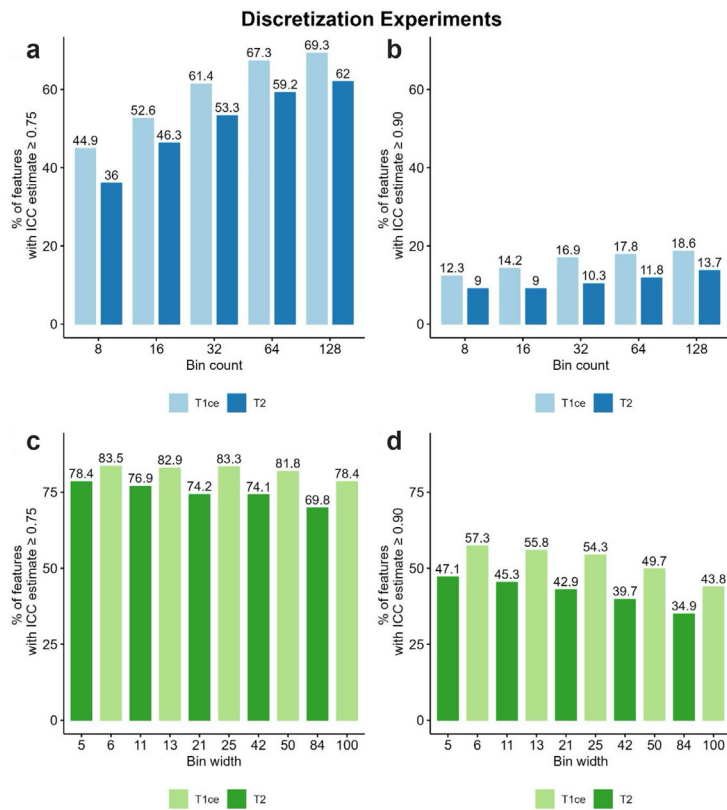
Figure 3. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility based on experiments with discretization parameters. Experiments with bin count **(a, b)** and bin width **(c, d)** on contrast-enhanced T1-weighted (T1ce) and T2-weighted (T2) sequences. ICC, intraclass correlation coefficient.

Figure 4. Percentage of reproducible features based on experiments on resampling parameters, using ICC cut-off values of 0.75 **(a)** and 0.90 **(b)** for good and excellent reproducibility, respectively. Experiments were performed on contrast-enhanced T1-weighted (T1ce) and T2-weighted (T2) sequences. ICC, intraclass correlation coefficient.

and supported each other. Conversely, the most recent guidelines released by the IBSI,[5] and a recent seminal phantom study,[8] recommend relative discretization techniques (i.e., the bin count method) across disparate acquisitions. Despite the recommendations, some other studies have shown that the relative discretization method might not be the optimal technique.[24]

Lu et al.[25] investigated the robustness of PET/CT-based radiomic features in terms of segmentation and discretization and conducted experiments to study them in patients with nasopharyngeal carcinomas. In total, 50%–63% of their features had an ICC ≥0.8 for the segmentation experiments, whereas 21%–23% of features showed an ICC ≥0.8 for the discretization experiments. However, only 6 of 57 features (11%) had an ICC ≥0.8 for the simultaneous evaluation of both segmentation and discretization experiments. Although Lu et al.[25] used a methodology that was quite different from ours, their study was indeed successful in showing the impact of discretization on the segmentation-based reproducibility of the radiomic features.

Unlike the above-mentioned studies, we additionally experimented with resampling parameters and discovered that increasing resampling size resulted in improved segmentation-based reproducibility rates. This additional finding on resampling is contradictory to the studies on the phantom experiments regarding the reproducibility of radiomic feature values. For instance, in a very recent phantom study, Wichtmann et al.[8] recommended that resampled voxels should not be too far from the original voxel size regarding feature reproducibility.

Our experiments and previous studies indicate that both discretization and resampling parameters significantly impact the segmentation-based reproducibility of radiomic features, and the optimal parameters to achieve high reproducibility in feature values and segmentation-based reproducibility seem contradictory. For this reason, care should be taken to find the optimal parameters to achieve both feature value reproducibility and segmentation-wise reproducible features within the radiomic pipeline.

This study has several differences when compared with previous studies. First, the number of features was higher than that of previous studies and was as high as those in radiomics research publications that had a clinical purpose. Second, the analysis was not limited to discretization but included ex-

(i.e., bin count method) discretization parameters and studied the distribution and highest number of replicable features for each technique. In addition, they utilized computer-generated delineations that were indicative of inter-observer variability. They observed that the discretization approach had a direct impact on feature repeatability, independent of observers, software, or method of delineation (simulated vs. human). Absolute discretization (i.e., the fixed bin width method) was recommended because it con-

sistently produced statistically considerably more reproducible features than relative discretization. Large bin numbers or narrow bin widths produced the highest number of repeatable features in all experiments. They also underlined that, regardless of the selected method, detailed documentation is vital so that results can be replicated. Although the tumors and range of parameters were completely different in our study from those of Duron et al.[24], we observed similar trends in discretization experiments that confirmed

Table 4. Comparison of reproducible features between different discretization techniques

ICC cut-off	Sequence	Bin count vs. bin width	Statistic[1]	df	P[2]
0.75	T1ce	128 vs. 6	104.4	1	<0.001
		16 vs. 50	270.7	1	<0.001
		32 vs. 25	182.6	1	<0.001
		64 vs. 13	114.2	1	<0.001
		8 vs. 100	293.5	1	<0.001
	T2	128 vs. 5	120.5	1	<0.001
		16 vs. 42	242.0	1	<0.001
		32 vs. 21	152.5	1	<0.001
		64 vs. 11	124.3	1	<0.001
		8 vs. 84	307.8	1	<0.001
0.90	T1ce	128 vs. 6	398.1	1	<0.001
		16 vs. 50	365.0	1	<0.001
		32 vs. 25	380.7	1	<0.001
		64 vs. 13	386.7	1	<0.001
		8 vs. 100	325.4	1	<0.001
	T2	128 vs. 5	322.7	1	<0.001
		16 vs. 42	321.8	1	<0.001
		32 vs. 21	338.4	1	<0.001
		64 vs. 11	337.0	1	<0.001
		8 vs. 84	300.0	1	<0.001

[1]McNemar's chi-squared. [2]In all comparisons, bin width was superior to bin count in terms of proportions of reproducible features. ICC, intraclass correlation coefficient; T1ce, contrast-enhanced T1-weighted; T2, T2-weighted; df, degrees of freedom.
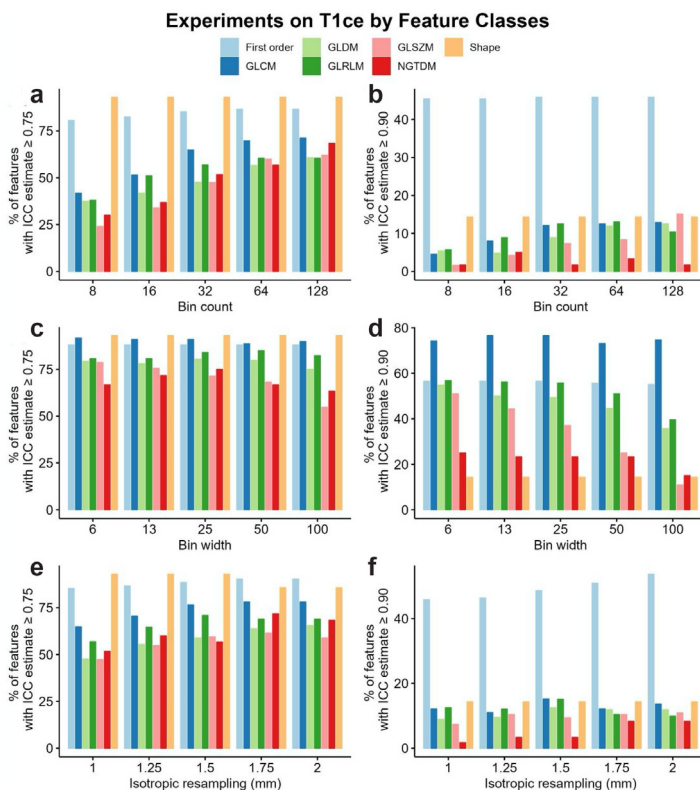
Figure 5. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility by feature classes. Experiments with bin count **(a, b)**, bin width **(c, d)**, and resampling **(e, f)** on contrast-enhanced T1-weighted (T1ce) sequence. ICC, intraclass correlation coefficient; GLCM, gray-level cooccurrence matrix; GLSZM, gray-level size zone matrix; GLRLM, gray-level run-length matrix; GLDM, gray-level dependence matrix; NGTDM, neighboring gray-tone difference matrix.

periments regarding resampling. These two preprocessing options commonly appear in open-source feature extraction software programs. Third, the experiments were performed in a different pathology (i.e., glioma), expanding the knowledge of the impact of preprocessing on segmentation-based reproducibility of radiomic features.

The public annotation dataset of BraTS 2021 was not used in the reproducibility experiments of this study because those data were based on a fusion of resultant annotations from several automated methods, first using the simultaneous truth and performance level estimation algorithm, followed by corrections applied by experts.[28] It would be difficult to perform and replicate the reproducibility experiments based on the public dataset, which may also not be representative of radiomics publications in general (not specifically those on gliomas) because those papers assessing segmentation reproducibility generally include at least two individual readers. For this reason, we segmented the dataset included in this study ourselves using the whole tumor volume to truly represent the segmentation-based reproducibility step of the radiomic studies.

Our experiments provided several practical points that might be considered in radiomic pipelines, associated publications, and clinical applications. First, image processing including discretization and voxel resampling has a considerable impact on the segmentation-based reproducibility of radiomic features; this should be considered as a means of improving the reproducibility of radiomic features that will be input to the following modeling stages. Second, the bin width method provided more reliable features than the bin count method in terms of segmentation-based reproducibility. Therefore, the bin width method should be favored in clinical studies. Third, using lower values for the bin width and higher values for the resampling provided more reproducible features. Given that there has been a lack of standardized preprocessing settings for discretization and resampling in the literature, these findings might provide guidance for end-users of the radiomic feature extraction tools. Fourth, due to their influence on the generation of reproducible inputs for modeling, our findings indicate that the preprocessing methods and their parameters must be defined in detail in published articles for radiomics models to be reliable.[35] According to a recent study, these essential radiomic parameters have been usually poorly reported in publications.[7] The recently published

Experiments on T1ce by Image Types
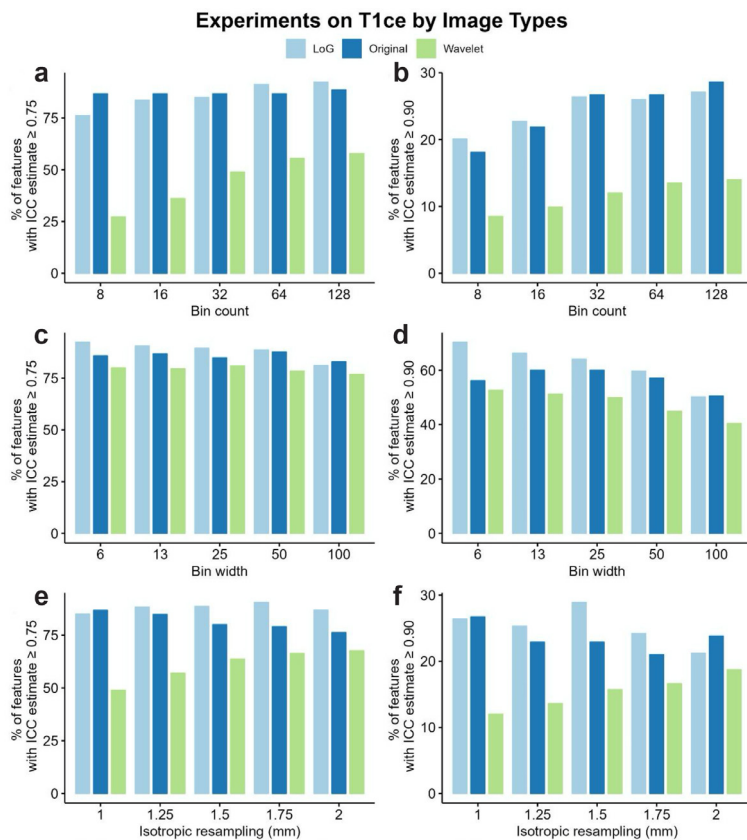
Figure 6. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility by image types. Experiments with bin count (**a, b**), bin width (**c, d**), and resampling (**e, f**) on contrast-enhanced T1-weighted (T1ce) sequence. ICC, intraclass correlation coefficient.

Checklist for Evaluation of Radiomics Research has also drawn attention to the same reporting issues.[9]

Our findings in this study should be interpreted with the following limitations.

First, the protocol for the acquisition of the BraTS 2021 challenge is not entirely clear. It is necessary to conduct research into the influence of the acquisition protocol (e.g., scanner type or acquisition settings) on image properties to gain a deeper comprehension of the behavior of radiomic features.

Second, our research was limited to a single imaging modality, two sequences, manual three-dimensional segmentation, a single tumor pathology, and gross tumor volume to remain manageable, considering the number of experiments conducted. However, we should acknowledge that every one of the aforementioned limitations may hamper the generalizability of the findings. We could also have added other alternatives to this study; however, that may have unnecessarily increased the complexity and workload, which was already high. This study aimed primarily to bring the attention of the radiomics community to the sensitivity of segmentation-based reproducibility to slight changes in two common preprocessing methods and offer reasonable settings. Alternative factors, such as different tumors, other MRI sequences, and different segmentation techniques, should be investigated as part of ongoing research.

Third, although significant and recommended by the IBSI guidelines,[5] the preprocessing techniques utilized in this study were only representative of a subset of the available options. However, the methods we used are available on the user interface of nearly all open-source radiomic feature extraction tools. The issue of standardization in radiomic studies may also involve scanner performance, acquisition protocols, acquisition sequence parameters, and data analysis techniques. However, we believe that the results of our study could be a step toward the standardization of radiomics.

Fourth, in our resampling experiments, the bin count was fixed. In light of the pairwise comparison experiments that were conducted with the final number of gray levels fixed, we anticipate observing a similar pattern when employing the bin width method. Additionally, when resampling images, we performed downsampling, as there has been no clear evidence on whether upsampling or downsampling methods are preferable.[2,5,8] However, although we considered the use of upsampling to be counterintuitive due to the addition of new voxels, it should be further explored in future experiments.

Fifth, the optimal settings for image processing to achieve the highest proportion of reproducible features were specific to the configuration used in this study. Our objective was not to identify absolute optimal values for all combinations of preprocessing settings. Consequently, no definitive conclusions should be drawn regarding the absolute best parameters (because, for example, they may be beyond the range of parameters used in the experiments) or the optimal sequence and discriminative performance.

Sixth, we did not test semi-automated or automated procedures in this study. Even with such techniques, a human touch or consensus segmentation is usually needed for correction, necessitating an analysis of feature reproducibility for segmentation, and supporting the need for conducting such a study.

In conclusion, to improve and standardize radiomic applications, every potential dependency of radiomic features on various parts of the radiomic workflow should be considered while developing a clinical or research project. In this study, the effect of image preprocessing parameters on the segmentation-based reproducibility of radiomic features from MRI was investigated systematically. Variations of image processing parameters related to discretization and resampling had a significant impact on the segmentation-based reproducibility of radiomic features within the scope of this study, regardless of MRI sequences. In terms of segmentation-based reproducibility, the bin width method yielded more reliable features than the bin count method. Using lower bin width values and higher resampling values produced more reproducible features. We recommend that these processing parameters be determined within the radiomic pipeline and transparently reported in radiomic publications. We anticipate that the implementation of our recommendations may facilitate the selection of more reproducible features and enhance the translation and generalizability of radiomics analyses. Considering the radiomics reproducibility crisis, extensive reproducibility studies are required before radiomics can be reliably implemented in routine clinical practice.

References

1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology*. 2016;278(2):563-577. [CrossRef]

2. van Timmeren JE, Cester D, Tanadini-Lang S, Alkadhi H, Baessler B. Radiomics in medical imaging-"how-to" guide and critical reflection. *Insights Imaging*. 2020;11(1):91. [CrossRef]

3. Rogers W, Thulasi Seetha S, Refaee TAG, et al. Radiomics: from qualitative to quantitative imaging. *Br J Radiol*. 2020;93(1108):20190948. [CrossRef]

4. Kocak B, Baessler B, Cuocolo R, Mercaldo N, Pinto Dos Santos D. Trends and statistics of artificial intelligence and radiomics research in radiology, nuclear medicine, and medical imaging: bibliometric analysis. *Eur Radiol*. 2023;33(11):7542-7555. [CrossRef]

5. Zwanenburg A, Vallières M, Abdalah MA, et al. The Image biomarker standardization initiative: standardized quantitative radiomics for high-throughput image-based phenotyping. *Radiology*. 2020;295(2):328-338. [CrossRef]

6. Hagiwara A, Fujita S, Ohno Y, Aoki S. Variability and standardization of quantitative imaging: monoparametric to multiparametric quantification, radiomics, and artificial intelligence. *Invest Radiol*. 2020;55(9):601-616. [CrossRef]

7. Kocak B, Yuzkan S, Mutlu S, Bulut E, Kavukoglu I. Publications poorly report the essential RadiOmics ParametERs (PROPER): a meta-research on quality of reporting. *Eur J Radiol*. 2023;167:111088. [CrossRef]

8. Wichtmann BD, Harder FN, Weiss K, et al. Influence of image processing on radiomic features from magnetic resonance imaging. *Invest Radiol*. 2023;58(3):199-208. [CrossRef]

9. Kocak B, Baessler B, Bakas S, et al. CheckList for EvaluAtion of Radiomics research (CLEAR): a step-by-step reporting guideline for authors and reviewers endorsed by ESR and EuSoMII. *Insights Imaging*. 2023;14(1):75. [CrossRef]

10. Peerlings J, Woodruff HC, Winfield JM, et al. Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial. *Sci Rep*. 2019;9(1):4800. [CrossRef]

11. van Timmeren JE, Leijenaar RTH, van Elmpt W, et al. Test-retest data for radiomics feature stability analysis: generalizable or study-specific? *Tomography*. 2016;2(4):361-365. [CrossRef]

12. Traverso A, Wee L, Dekker A, Gillies R. Repeatability and reproducibility of radiomic features: a systematic review. *Int J Radiat Oncol Biol Phys*. 2018;102(4):1143-1158. [CrossRef]

13. Zhovannik I, Bussink J, Traverso A, et al. Learning from scanners: bias reduction and feature correction in radiomics. *Clin Transl Radiat Oncol*. 2019;19:33-38. [CrossRef]

14. Mackin D, Fave X, Zhang L, et al. Measuring computed tomography scanner variability of radiomics features. *Invest Radiol*. 2015;50(11):757-765. [CrossRef]

15. Gitto S, Cuocolo R, Emili I, et al. Effects of interobserver variability on 2D and 3D CT- and MRI-based texture feature reproducibility of cartilaginous bone tumors. *J Digit Imaging*. 2021;34(4):820-832. [CrossRef]

16. Kocak B, Ates E, Durmaz ES, Ulusan MB, Kilickesmez O. Influence of segmentation margin on machine learning-based high-dimensional quantitative CT texture analysis: a reproducibility study on renal clear cell carcinomas. *Eur Radiol*. 2019;29(9):4765-4775. [CrossRef]

17. Zhao B, Tan Y, Tsai WY, et al. Reproducibility of radiomics for deciphering tumor phenotype with imaging. *Sci Rep*. 2016;6:23428. [CrossRef]

18. Alis D, Yergin M, Asmakutlu O, Topel C, Karaarslan E. The influence of cardiac motion on radiomics features: radiomics features of non-enhanced CMR cine images greatly vary through the cardiac cycle. *Eur Radiol*. 2021;31(5):2706-2715. [CrossRef]

19. Fornacon-Wood I, Mistry H, Ackermann CJ, et al. Reliability and prognostic value of radiomic features are highly dependent on choice of feature extraction platform. *Eur Radiol*. 2020;30(11):6241-6250. [CrossRef]

20. Ubaldi L, Saponaro S, Giuliano A, Talamonti C, Retico A. Deriving quantitative information from multiparametric MRI via radiomics: evaluation of the robustness and predictive value of radiomic features in the discrimination of low-grade versus high-grade gliomas with machine learning. *Phys Med*. 2023;107:102538. [CrossRef]

21. Park JE, Park SY, Kim HJ, Kim HS. Reproducibility and generalizability in radiomics modeling: possible strategies in radiologic and statistical perspectives. *Korean J Radiol*. 2019;20(7):1124-1137. [CrossRef]

22. Zhao B. Understanding sources of variation to improve the reproducibility of radiomics. *Front Oncol*. 2021;11:633176. [CrossRef]

23. Koçak B, Durmaz EŞ, Ateş E, Kılıçkesmez Ö. Radiomics with artificial intelligence: a practical guide for beginners. *Diagn Interv Radiol*. 2019;25(6):485-495. [CrossRef]

24. Duron L, Balvay D, Vande Perre S, et al. Gray-level discretization impacts reproducible MRI radiomics texture features. *PLoS One*. 2019;14(3):e0213459. [CrossRef]

25. Lu L, Lv W, Jiang J, et al. Robustness of radiomic features in [^{11}C]Choline and [^{18}F]FDG PET/CT imaging of nasopharyngeal carcinoma: impact of segmentation and discretization. *Mol Imaging Biol*. 2016;18(6):935-945. [CrossRef]

26. Bakas S, Akbari H, Sotiras A, et al. Advancing the cancer genome atlas glioma MRI collections with expert segmentation labels and radiomic features. *Sci Data*. 2017;4:170117. [CrossRef]

27. Menze BH, Jakab A, Bauer S, et al. The Multimodal Brain Tumor Image Segmentation Benchmark (BRATS). *IEEE Trans Med Imaging*. 2015;34(10):1993-2024. [CrossRef]

28. Baid U, Ghodasara S, Mohan S, et al. The RSNA-ASNR-MICCAI BraTS 2021 Benchmark on Brain Tumor Segmentation and Radiogenomic Classification. Published online September 12, 2021. [CrossRef]

29. Rohlfing T, Zahr NM, Sullivan EV, Pfefferbaum A. The SRI24 multichannel atlas of normal adult human brain structure. *Hum Brain Mapp*. 2010;31(5):798-819. [CrossRef]

30. Thakur S, Doshi J, Pati S, et al. Brain extraction on MRI scans in presence of diffuse glioma: multi-institutional performance evaluation of deep learning methods and robust modality-agnostic training. *Neuroimage*. 2020;220:117081. [CrossRef]

31. van Griethuysen JJM, Fedorov A, Parmar C, et al. Computational radiomics system to decode the radiographic phenotype. *Cancer Res*. 2017;77(21):e104-e107. [CrossRef]

32. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. *Psychol Bull*. 1979;86(2):420-428. [CrossRef]

33. Koo TK, Li MY. A Guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155-163. [CrossRef]

34. Traverso A, Kazmierski M, Shi Z, et al. Stability of radiomic features of apparent diffusion coefficient (ADC) maps for locally advanced rectal cancer in response to image pre-processing. *Phys Med*. 2019;61:44-51. [CrossRef]

35. Carré A, Klausner G, Edjlali M, et al. Standardization of brain MR images across machines and protocols: bridging the gap for MRI-based radiomics. *Sci Rep*. 2020;10(1):12340. [CrossRef]

Supplementary Table S1. Patient identifiers

#	Identifier	#	Identifier
1	BraTS2021_00134	26	BraTS2021_01028
2	BraTS2021_00138	27	BraTS2021_01089
3	BraTS2021_00147	28	BraTS2021_01161
4	BraTS2021_00167	29	BraTS2021_01180
5	BraTS2021_00221	30	BraTS2021_01251
6	BraTS2021_00233	31	BraTS2021_01254
7	BraTS2021_00247	32	BraTS2021_01302
8	BraTS2021_00271	33	BraTS2021_01357
9	BraTS2021_00306	34	BraTS2021_01359
10	BraTS2021_00316	35	BraTS2021_01360
11	BraTS2021_00317	36	BraTS2021_01365
12	BraTS2021_00364	37	BraTS2021_01380
13	BraTS2021_00373	38	BraTS2021_01426
14	BraTS2021_00446	39	BraTS2021_01447
15	BraTS2021_00453	40	BraTS2021_01465
16	BraTS2021_00514	41	BraTS2021_01476
17	BraTS2021_00557	42	BraTS2021_01479
18	BraTS2021_00575	43	BraTS2021_01491
19	BraTS2021_00577	44	BraTS2021_01537
20	BraTS2021_00612	45	BraTS2021_01578
21	BraTS2021_00744	46	BraTS2021_01585
22	BraTS2021_00758	47	BraTS2021_01610
23	BraTS2021_00836	48	BraTS2021_01613
24	BraTS2021_01000	49	BraTS2021_01614
25	BraTS2021_01003	50	BraTS2021_01626

Experiments on T2 by Feature Classes
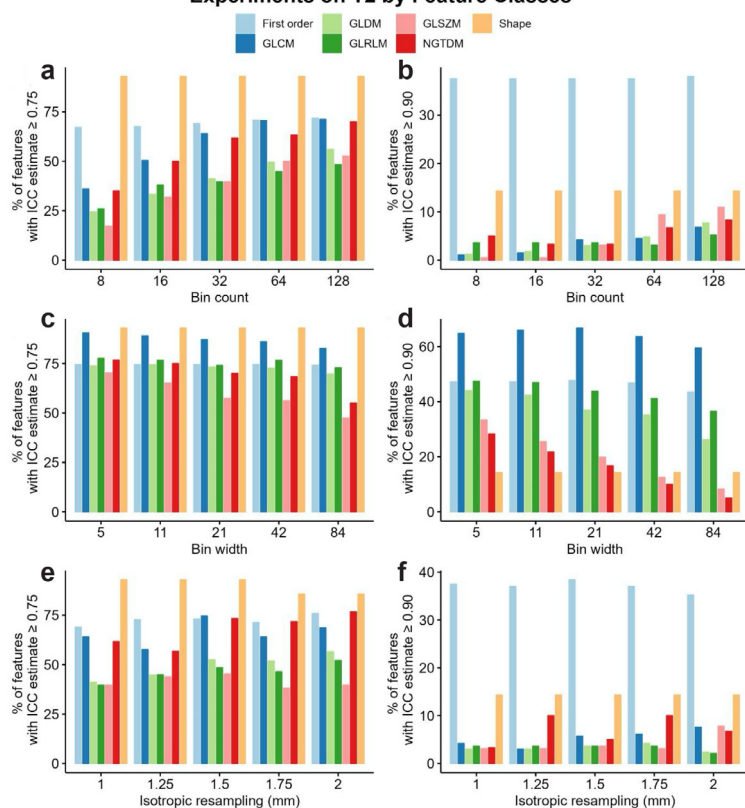
Supplementary Figures S1. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility by radiomic feature classes. Experiments with bin count **(a, b)**, bin width **(c, d)**, and resampling **(e, f)** on T2-weighted (T2) sequence. ICC, intraclass correlation coefficient; GLCM, gray-level cooccurrence matrix; GLSZM, gray-level size zone matrix; GLRLM, gray-level run-length matrix; GLDM, gray-level dependence matrix; NGTDM, neighboring gray-tone difference matrix.

Experiments on T2 by Image Types
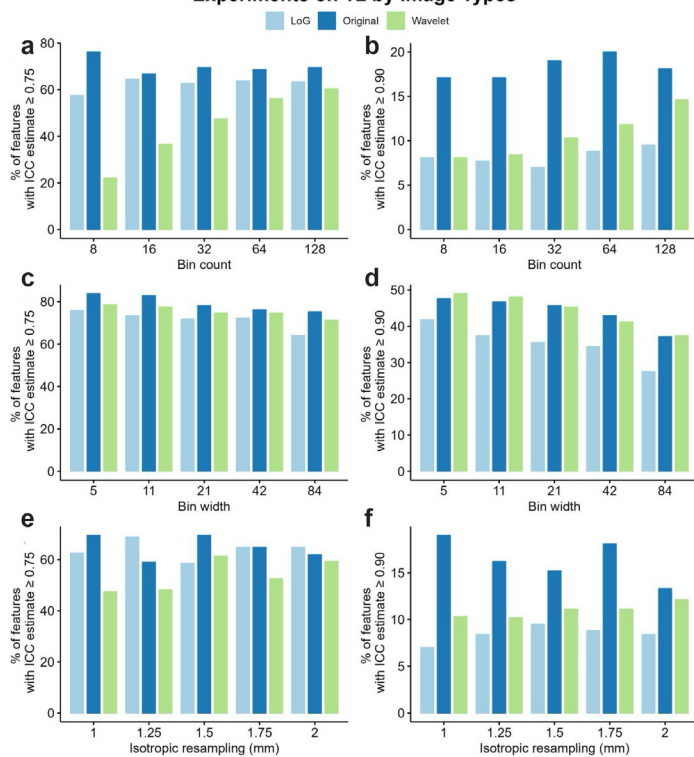
Supplementary Figures S2. Percentage of features with good (ICC ≥0.75) and excellent (ICC ≥0.90) reproducibility by image types. Experiments with bin count **(a, b)**, bin width **(c, d)**, and resampling **(e, f)** on T2-weighted (T2) sequence. ICC, intraclass correlation coefficient.

Educating the next generation of radiologists: a comparative report of ChatGPT and e-learning resources

İsmail Meşe[1]
Ceylan Altıntaş Taşlıçay[2]
Beyza Nur Kuzan[3]
Taha Yusuf Kuzan[4]
Ali Kemal Sivrioğlu[5]

[1]University of Health Sciences Türkiye, Erenköy Mental Health and Neurology Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

[2]MD Anderson Cancer Center, Department of Radiology, Houston, Texas, United States

[3]Kartal Dr. Lütfi Kırdar City Hospital, Clinic of Radiology, İstanbul, Türkiye

[4]Sancaktepe Şehit Prof. Dr. İlhan Varank Training and Research Hospital, Clinic of Radiology, İstanbul, Türkiye

[5]Liv Hospital Vadistanbul, Clinic of Radiology, İstanbul, Türkiye

ABSTRACT

Rapid technological advances have transformed medical education, particularly in radiology, which depends on advanced imaging and visual data. Traditional electronic learning (e-learning) platforms have long served as a cornerstone in radiology education, offering rich visual content, interactive sessions, and peer-reviewed materials. They excel in teaching intricate concepts and techniques that necessitate visual aids, such as image interpretation and procedural demonstrations. However, Chat Generative Pre-Trained Transformer (ChatGPT), an artificial intelligence (AI)-powered language model, has made its mark in radiology education. It can generate learning assessments, create lesson plans, act as a round-the-clock virtual tutor, enhance critical thinking, translate materials for broader accessibility, summarize vast amounts of information, and provide real-time feedback for any subject, including radiology. Concerns have arisen regarding ChatGPT's data accuracy, currency, and potential biases, especially in specialized fields such as radiology. However, the quality, accessibility, and currency of e-learning content can also be imperfect. To enhance the educational journey for radiology residents, the integration of ChatGPT with expert-curated e-learning resources is imperative for ensuring accuracy and reliability and addressing ethical concerns. While AI is unlikely to entirely supplant traditional radiology study methods, the synergistic combination of AI with traditional e-learning can create a holistic educational experience.

KEYWORDS

Artificial intelligence, ChatGPT, digital case studies, educational videos, radiology education

Corresponding author: İsmail Meşe

E-mail: ismail_mese@yahoo.com

Rapid advances in technology have revolutionized medical education, changing the way healthcare professionals share, access, and assimilate information.[1] This change is particularly evident in radiology, which relies heavily on advanced imaging techniques and visual data.[2] Continuous technological advancement has led to a significant change in approach, altering traditional teaching methods and introducing new resources that are transforming the educational environment.[3]

Electronic learning (e-learning) resources, including digital case studies and educational videos, are now essential sources of information.[4-6] Radiology, as a primarily visual discipline, has benefited significantly from these multimedia resources.[7] E-learning platforms, with their ability to present complex details through high-quality images, animations, and videos, provide radiologists with a powerful and thorough learning experience.[4,8] These digital resources also promote self-directed learning, giving professionals the flexibility to manage their educational journey based on their personal and professional commitments.[9] This independence promotes improved comprehension and retention, establishing e-learning as a fundamental aspect of modern radiology education. Table 1 summarizes common radiology e-learning resources and their characteristics.

The incorporation of artificial intelligence (AI)-powered tools, such as Chat Generative Pre-trained Transformer (ChatGPT), has introduced a dynamic, stimulating, and customized method of radiology education. OpenAI's ChatGPT is an ideal example of how chatbot technology can create tailored and effortless learning journeys.[10] Unlike standardized learning resourc-

es, these chatbots adapt to individual users, providing responses and guidance in real time, mimicking the experience of a personal tutor.[11] This provides learners with an unprecedented combination of immediacy and personalization, breaking down complex radiological ideas into manageable chunks of information and encouraging active participation. In the current literature, there is a noticeable gap in comparing AI tools, such as ChatGPT, directly with traditional e-learning platforms for radiology education. The lack of a direct comparison is significant because it misses an opportunity to evaluate their respective educational impacts, potential synergies, and the ways they could be tailored to improve radiologists' learning experiences. The existing focus on separate evaluations of e-learning platforms and AI tools without a comprehensive comparison within radiology education indicates a critical void in the literature, which may lead to biases in understanding the full potential and shortcomings of each educational technology in this specialized field.[8,12-14] The purpose of this review is to analyze the ChatGPT tool and compare it with e-learning resources for medical education, with a specific focus on radiology. Through an exhaustive review of empirical studies, surveys, and expert opinions, this paper elucidates the relative strengths and weaknesses of these tools in areas such as personalization, interactivity, visual learning,

complex concept presentation, accessibility, user experience, learning curves, and cost-effectiveness. This analysis is particularly relevant, as the global shift to distance learning, accelerated by the coronavirus disease-2019 (COVID-19) outbreak, requires a reassessment of educational tools.

Methodology

A comprehensive search was undertaken in traditional medical databases to discern relevant articles focusing on the use of e-learning platforms and ChatGPT in radiology education (Figure 1). The databases inspected include PubMed, Scopus, and Web of Science. Additionally, preprint servers, such as arXiv, bioRxiv, and medRxiv, were explored to capture the most recent and evolving literature on the subject. The search, conducted between January 2023 and October 2023, utilized keywords such as "e-learning in radiology," "ChatGPT in radiology," "ChatGPT in education," "AI-tools in radiology," and "AI-tools in education." The inclusion criteria were articles and preprints that investigated the deployment of e-learning platforms in radiology education, highlighting their advantages, disadvantages, and inherent features, as well as those that examined the application of ChatGPT in radiology teaching, emphasizing its potential, merits, and demerits.

The exclusion criteria encompassed articles that did not address e-learning platforms or ChatGPT in the context of radiology education and those that were duplicates of previously recognized articles. Each study was independently evaluated by two reviewers in terms of the inclusion and exclusion criteria, with any discrepancies resolved through discussion or with a third reviewer if necessary. Throughout the review procedure, the authors collaborated intensively. They teamed up for numerous tasks, including article selection, quality assessment, and the consolidation of findings. This cooperative methodology enables a thorough examination of the articles and ensures that the final review encapsulates a mutual agreement among the researchers. The data were synthesized from the selected articles using narrative methods by organizing them to identify common themes and divergent views on the use of e-learning platforms and ChatGPT in radiology education. These were presented descriptively and structured around the advantages, disadvantages, and educational outcomes reported by the studies.

E-learning resources in radiology education

The introduction of the internet has had a profound impact on various industries,

Main points

- Chat Generative Pre-Trained Transformer (ChatGPT) utilizes machine learning algorithms to provide a distinctive form of educational engagement. This allows for immediate and personalized feedback based on individual learner needs. Electronic learning (e-learning) platforms, despite their interactivity, may not offer the same level of real-time personalization.

- E-learning tools hold an advantage in teaching visually complex concepts. These tools may require fees and may not be as up-to-date. However, costs associated with these tools can potentially be reduced with ChatGPT, providing an accessible artificial intelligence (AI)-driven learning experience.

- Utilizing ChatGPT in radiology education presents challenges, such as data dependency and potential biases. To ensure precise and unbiased instruction, it is imperative to integrate ChatGPT with expert-curated resources, peer-reviewed content, and practical case studies. Collaboration between AI developers and medical specialists is essential to uphold the integrity, accuracy, and ethical standards of information dissemination.

Table 1. Websites related to radiology education, including video resources name description

Radiopaedia (https://radiopaedia.org/)	A collaborative radiology resource offering a vast collection of educational articles, cases, and quizzes
AuntMinnie (https://www.auntminnie.com/)	A comprehensive radiology website featuring news, case studies, webinars, and forums for radiology professionals
Radiology Assistant (https://radiologyassistant.nl/)	An educational resource for radiology residents, featuring illustrated articles, case studies, and quizzes
LearningRadiology (http://www.learningradiology.com/)	Provides radiology lectures, podcasts, and case studies for medical students, residents, and technologists
MedCram - Radiology (https://www.youtube.com/@Medcram)	Offers a collection of video lectures and courses on radiology topics, aimed at medical students, residents, and healthcare professionals
RadiologyEducation.com (https://radiologyeducation.com/)	A digital library of radiology resources, including textbooks, journals, and teaching files
Radiology Video Lectures (https://www.youtube.com/@RadiologyVideo)	A YouTube channel with video lectures on various radiology topics, aimed at medical students and radiology residents
Radiology Channel (https://www.youtube.com/@RadiologyChannel)	A YouTube channel that offers a series of video lectures, case studies, and tutorials related to radiology
Imaios (https://www.imaios.com/en)	A platform that provides e-learning courses in radiology and other medical disciplines
FreitasRad (https://www.freitasrad.net/)	A radiology education website that offers various learning resources, including articles, webinars, and case studies
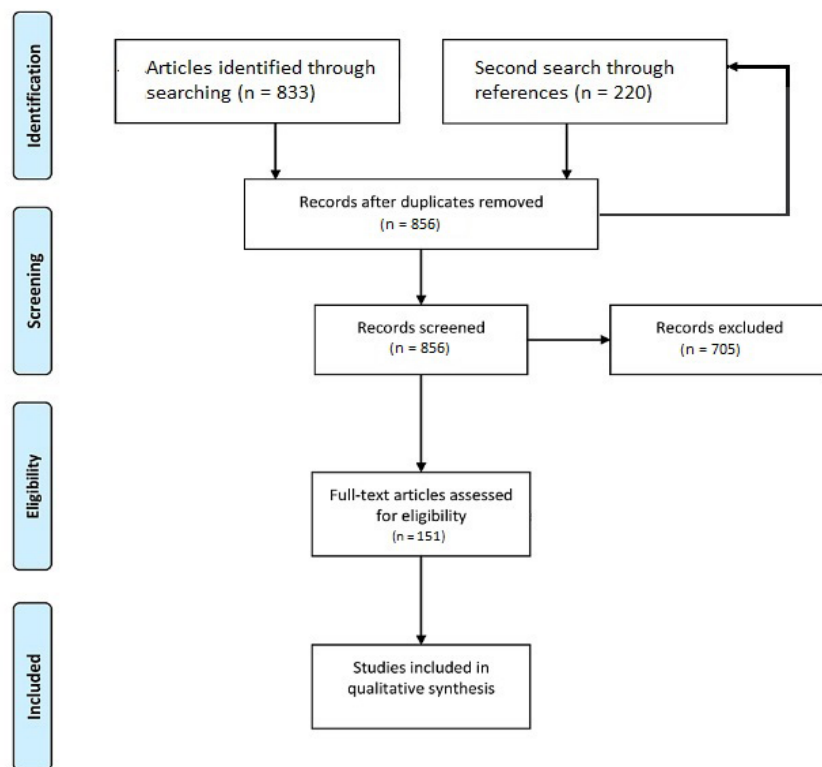
Figure 1. Selection process of articles used in this study. Adopted from the Preferred Reporting Items for Systematic Reviews and Meta-Analyses.

including medical education. Radiology education, in particular, has undergone significant transformation due to digital advancements at all levels of medical training.[15] This transformation goes beyond a mere change in format and addresses critical limitations inherent in traditional teaching methods. For example, e-learning resources overcome geographical and time limitations, presenting a broader range of case studies for examination.[16]

Traditionally, the teaching file has been the foundation of radiology education, containing a carefully curated set of cases that emphasize important diagnostic features. The emergence of the internet has elevated this concept by introducing online databases of expert-reviewed radiological images that function as lively, globally accessible learning resources.[16] Researchers such as Kahn[16] have advanced the digital transition by creating internet-based collections of meticulously reviewed radiological images. These resources possess dual functionality, acting as educational instruments and tools for clinical decision-making. Detailed indexing, such as MeSH codes and patient information, allows for more precise searches compared with traditional systems, thus broadening educational access.[16-18]

Different e-learning platforms provide diverse experiences (Table 1). Some learning platforms in radiology are "educator-centric," using structured curricula that are especially advantageous in subspecialties, such as neuroradiology and pediatric radiology.[19] By contrast, other platforms adopt a more decentralized approach, supporting peer-to-peer learning but requiring strict quality control to prevent unreliable content.[20,21] The choice of a particular platform often depends on the learner's objectives and the specific learning context. In an intervention study by Salajegheh et al.[22], the research findings highlight the significant potential of e-learning to strengthen the improvement of X-ray interpretation skills among medical students. This study underscores the crucial importance of well-designed, interactive e-learning materials. Regardless of the chosen system, maintaining ongoing quality assurance and content curation is imperative due to the critical nature of radiological diagnosis and treatment.[19]

Considerations of accessibility and cost-effectiveness are essential in the selection and usage of e-learning resources. Despite the exceptional benefits of these digital platforms in content and flexibility, the need for high bandwidth and sometimes specialized hardware can restrict their

use, particularly in low-resource settings.[23] Hussein[23] investigated these factors by analyzing three variables: acceptance, accessibility, and savings. The research discovered that acceptance was the most important aspect determining the efficacy of e-learning, with accessibility and cost savings having a less significant impact.

Progressive technologies have the potential to enhance radiology e-learning experiences. Interactive functions, such as keypads and wireless interfaces, are frequently included, augmenting the learning experience.[20,21] Simulators provide immersive environments for practicing interventional techniques and comprehending hospital-based systems, such as Picture Archiving. Overdyk and McEvoy[24] evaluated the potential of handheld computers to enhance workflow efficiency, and the results indicated that the system has potential for educational purposes.[22]

The ethical challenges of online platforms in education include issues of privacy, surveillance, autonomy, bias, and discrimination.[25] In education, there are concerns about data breaches by tech companies and users inadvertently sharing excessive personal metadata.[25] There is a moral dilemma when students and parents are compelled to use online platforms or AI systems despite these privacy concerns in schools.[25] The surveillance capacity of these platforms threatens students' participation and sense of security, whereas predictive algorithms can undermine individual autonomy, possibly perpetuating societal biases.[25]

ChatGPT in radiology education

The rising role of ChatGPT in radiology education

ChatGPT, an AI-based language model developed by OpenAI, is becoming an increasingly valuable resource in radiology education.[26,27] Designed to understand and process natural language, ChatGPT can answer questions, provide explanations, and even suggest courses of action across a wide range of topics. Its conversational abilities, similar to those of a human, and its contextually relevant responses make it an effective tool for radiology residents and professionals.[27,28] ChatGPT is an advanced language model that uses deep learning techniques to produce human-like responses to natural language inputs. Built on the generative pre-trained transformer architecture, these models harness the power of deep learning to generate human-like text based on

patterns identified from vast amounts of textual data. A significant strength of ChatGPT is its expertise in natural language processing (NLP), enabling it to understand and produce coherent text. This proficiency, derived from comprehensive training on diverse datasets, allows it to simulate dialogue-like interactions pertinent to radiology or other subjects.[29]

NLP is the technology behind how computers understand and respond to human language, designed to bridge the gap between human communication and computer understanding. It uses a variety of techniques to interpret human language: tokenization, which breaks up a sentence into individual words or phrases called "tokens;" parsing, akin to diagramming a sentence, establishes the structure and word relationships within a sentence; sentiment analysis, which determines the emotional tone behind words; named entity recognition, involving identifying and classifying names of people, organizations, and locations into predefined categories; part-of-speech tagging, which assigns grammatical categories to each word, such as nouns, verbs, and adjectives; and machine translation, entailing text translation from one language to another, a complex task that requires understanding the grammar and context in both languages.[30]

Language modeling and generation underpin ChatGPT's text creation capabilities. The model is trained to predict the next word in a sequence, enabling the crafting of coherent sentences and paragraphs. When processing a prompt, ChatGPT first breaks down the input text into tokens. It then uses these tokens to grasp the context and meaning. To generate a response, ChatGPT predicts the next token in the sequence, based on the previous ones, and continues this prediction process until a complete answer is formed. Attention mechanisms are employed throughout, allowing the model to focus on different parts of the input as needed, which is essential for understanding context and generating pertinent responses. By integrating these NLP techniques, models such as ChatGPT can perform a variety of language understanding and production tasks, from conversational interactions to language translation and beyond.[30]

The use of AI and AI chatbots alongside traditional teaching methods has shown significant potential and has garnered substantial attention in research.[27,28,31] One of the primary advantages that ChatGPT brings is in the creation of learning assessments.[27,32] ChatGPT has demonstrated capability in generating exercises, quizzes, and scenarios that can be employed in a classroom setting to aid in resident practice and assessment (Figure 2). ChatGPT can streamline the design of learning assessments, potentially improving question quality. Zhai[33] suggests that educators can utilize ChatGPT to devise assessment items, enhancing the quality by adhering to standardized frameworks. How-

ever, it is important to note that although AI can suggest assessment tasks, these suggestions may not always encapsulate all the targeted learning objectives.[34] Hence, while Han et al.[35] utilized ChatGPT to draft a multiple-choice question for a medical topic, they also highlighted the importance of refining such AI-generated questions to better fit specific course requirements.[36]

Another advantage of pedagogical practices is the potential for educators to integrate innovative strategies, such as the flipped classroom approach, using ChatGPT.[37] ChatGPT's vast array of features has been highlighted by experts who emphasized its capabilities in generating comprehensive lesson plans and presentations (Figure 3).[37] Moreover, ChatGPT's potential as a virtual personal tutor is undeniable. Unlike traditional tutoring methods constrained by time and location, ChatGPT offers round-the-clock support.[27] Furthermore, AI tools provide feedback and tailored answers specific to students' academic queries.[28,35,38]

The ability to distill vast amounts of information is crucial for learners, and ChatGPT excels at this. It can process and summarize information efficiently, presenting it in an understandable manner, which can be particularly valuable for revision or to get a grasp on new topics (Figure 4).[39] Collaborative learning experiences are enriched through scenarios and structured discussions. ChatGPT has been successfully used to create dialogues for educational purposes (Figure 5). Specifically, when tasked with creating dialogues to be imported into Google Dialogflow, a popular platform for creating chatbots, ChatGPT succeeded in the task.[35] ChatGPT can generate varied scenarios for group activities, thereby aiding residents in their collaborative efforts. It can also lay down discussion structures and provide real-time feedback, thereby enhancing group discussions and debates.[40] This collaborative angle is emphasized further by Gilson et al.[41], who noted that the discourse in small-group problem-solving is particularly beneficial for student learning.

The advantages of tools such as ChatGPT have been highlighted for enhancing student outcomes and promoting critical thinking. In a 2023 study, Sánchez et al.[42] noted that ChatGPT promotes the development of critical thinking, particularly in a theoretical context. By promptly and accurately answering specific inquiries, ChatGPT provides students with relevant and timely information, enhancing their exploration and understanding of various subjects.
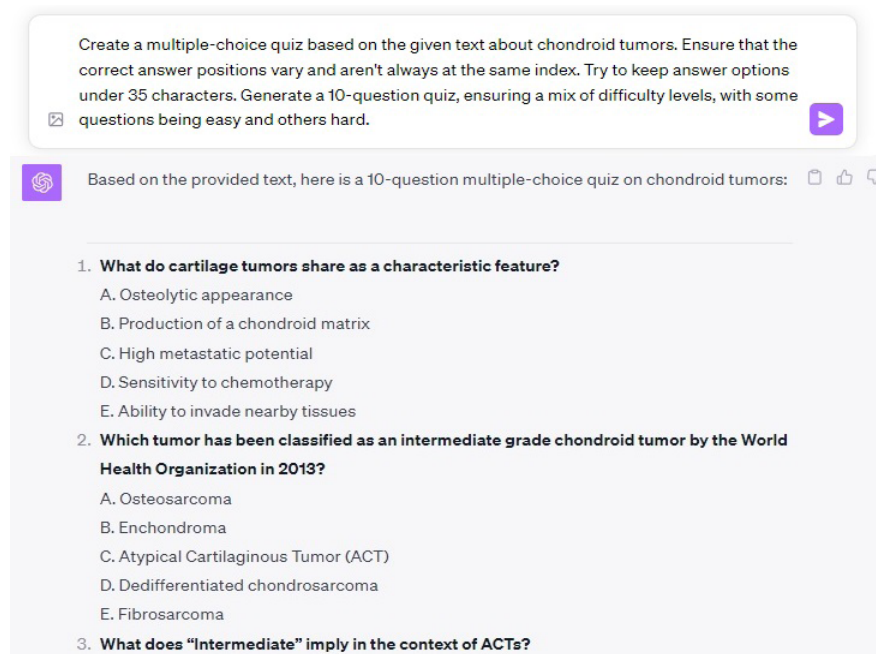
Figure 2. ChatGPT learning assessment creation. ChatGPT, Chat Generative Pre-Trained Transformer.
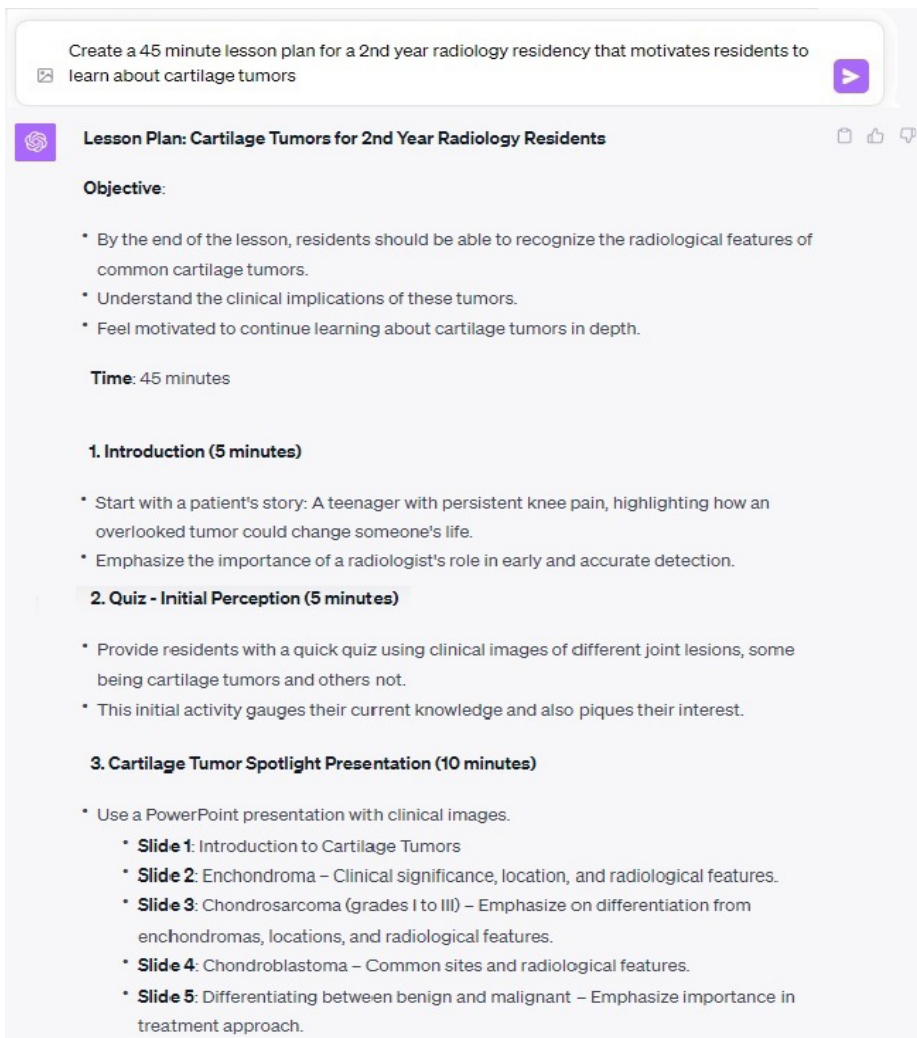
Figure 3. ChatGPT's lesson plan generation capabilities. ChatGPT, Chat Generative Pre-Trained Transformer.
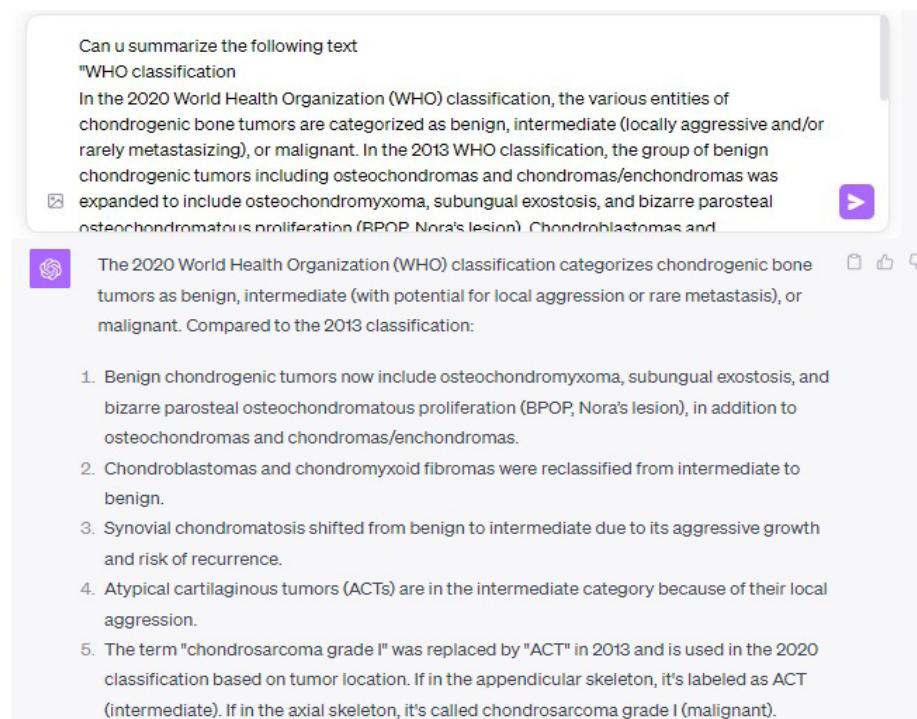
Figure 4. ChatGPT simplifies information. ChatGPT, Chat Generative Pre-Trained Transformer.

ChatGPT is particularly useful in preparing learners for high-stakes assessments, such as radiology board exams. By simulating diverse exam scenarios, including interactive question and answer sessions, ChatGPT aids in comprehensive exam preparation, allowing students to effectively enhance their knowledge and critical thinking skills.[43,44] AI can also be used for essay grading, though its effectiveness is still debated.[45] In addition, ChatGPT can be useful for translating educational materials into various languages, enhancing the accessibility of resources for a global student body (Figure 6).[35] Machine translation is not new, but AI models such as GPT-4 have continued to advance the field.[45]

A study by Ausat et al.[26] on the use of ChatGPT in learning emphasized that technology serves only as a tool and cannot fully replace the role of a teacher. The authors stress the importance of integrating technology appropriately and effectively and developing competence among teachers in managing learning with technology.

Tlili et al.[46] conducted a case study on ChatGPT's role in education. They stressed the importance of accurate content, noting inconsistencies in answers given to the same questions by different users. The team expressed concerns about the quality of assessments generated by ChatGPT and emphasized the need for clearer guidelines. While recognizing ChatGPT's potential, they highlighted its shortcomings, such as a lack of emotional depth and trustworthiness in responses. They underscored the need for careful inclusion of ChatGPT in educational contexts.[46]

The role of ChatGPT in radiology and academic applications

ChatGPT's utilization of NLP capacities assists radiologists in the analysis of medical images, presenting a significant advancement in radiology. When provided with information from an imaging modality's findings, the model produces a series of potential diagnoses for consideration (Figure 7).[14] To maximize ChatGPT's precision, it is crucial to supply it with specific prompts tailored to the medical images in question.[47] By providing detailed prompts that include the patient's medical history, symptoms, and distinctive image features, ChatGPT assists learners in enhancing their diagnostic skills. Additionally, ChatGPT is useful for giving real-time feedback during questioning, doubt clarification, or case discussions, thereby enriching the learning process.[28]

In the field of radiology, ChatGPT's utility is notably constrained by the discipline's inherent visual emphasis.[48] Although ChatGPT can provide assistance based on textual descriptions of findings, it may overlook errors introduced by human interpreters. In clinical scenarios, the need to describe an image in text to an AI, rather than allowing the AI to analyze it directly using pattern recognition algorithms, introduces an unnecessary step that may compromise efficiency. For radiological image analysis, specific AI technologies, including radiomics[49] and deep learning models,[50] are particularly suitable. These models, trained on extensive datasets, excel at identifying complex patterns in radiological images and extracting quantitative features, thereby enhancing their potential efficacy for diagnostic and predictive tasks. Building on the limitations of ChatGPT in radiology, the emergence of the Large Language-and-Vision Assistant (LLaVA) addresses some of these challenges.[13] As a visual chatbot capable of processing both textual and visual data, LLaVA offers a more integrated approach to AI in healthcare.[13] Its potential in radiology promises increased diagnostic accuracy and lighter workloads for radiologists.[29] It is crucial to understand that these AI instruments are designed to augment, not replace, human skills in radiology.

Figure 5. ChatGPT in educational dialogue creation. ChatGPT, Chat Generative Pre-Trained Transformer.

Figure 6. Translating educational materials into multiple languages with ChatGPT. ChatGPT, Chat Generative Pre-Trained Transformer.

With the latest updates, ChatGPT has gained the ability to interpret images. In a recent study, researchers evaluated GPT-4V(ision), an enhanced multimodal model that combines language and visual understanding. The study aimed to assess its capabilities across various tasks, input types, and operational modes, and to develop effective prompting strategies. The preliminary findings indicate that although GPT-4V is adept at managing multimodal inputs and shows promise in fostering novel interactive capabilities, such as visual referring prompts, it is not yet highly accurate and sometimes makes erroneous claims. The study suggests that GPT-4V could lead to more research into multimodal applications and enhanced problem-solving methods.[51] However, there is a lack of sufficient research or experience in interpreting radiological images. By understanding the capabilities and current limitations of GPT-4V in image interpretation, educators and professionals in radiology can begin to explore the integration of AI-assisted learning and diagnostic strategies. This emerging technology may revolutionize how radiological data are taught and interpreted, prompting a re-evaluation of curricula.

ChatGPT also excels as a research tool. It is notable for its ability to organize ideas for writing tasks. By providing it with prompts, ChatGPT can quickly generate outlines and assist in the initial stages of research and writing (**Figure** 8).[52] However, researchers should approach its output with caution and are advised to adjust and verify ideas produced by the tool to prevent inaccuracies.[37] ChatGPT plays an instrumental role in aiding students and researchers by helping them structure their thoughts and con-

cepts efficiently for writing assignments.[36,52] A study by Pech-Rodríguez et al.[53] evaluated ChatGPT's performance across various academic tasks. The findings indicated that the platform can generate grammatically correct essays; however, it often includes redundant information and may omit deeper insights. When tackling mathematical problems, ChatGPT showed inconsistencies, especially when its results were compared with those from MATLAB, suggesting potential reliability issues in academic tutoring. Moreover, although ChatGPT proved informative on general topics, it occasionally fell short in providing comprehensive answers to more intricate subjects. The study underscored the risks of overly depending on such platforms and highlighted the challenges in regulating this emerging technology.
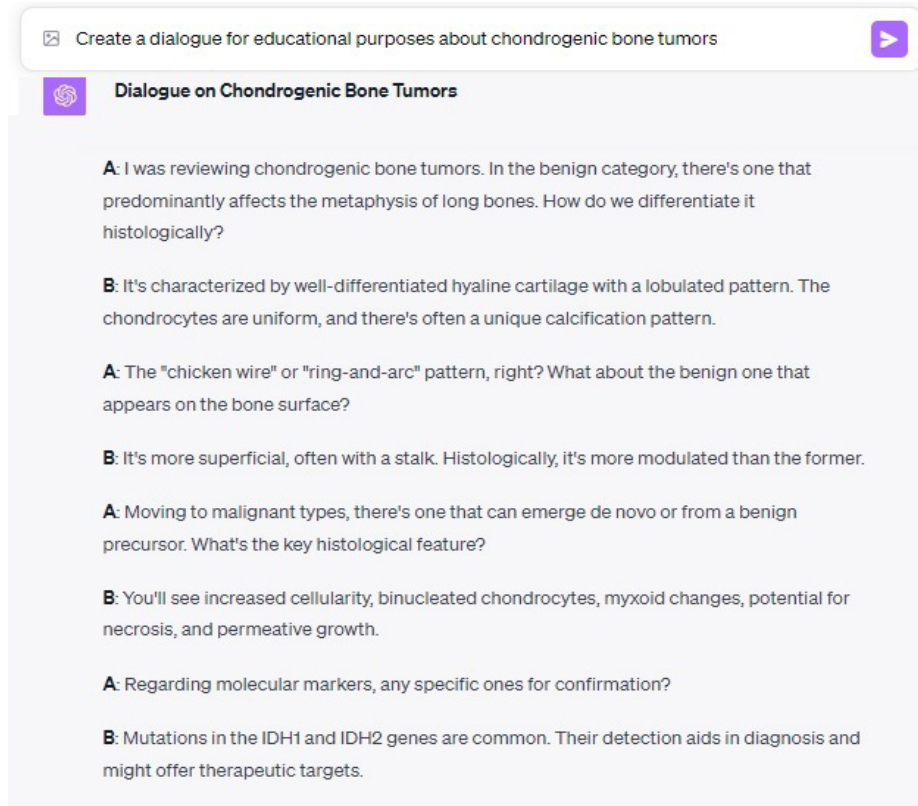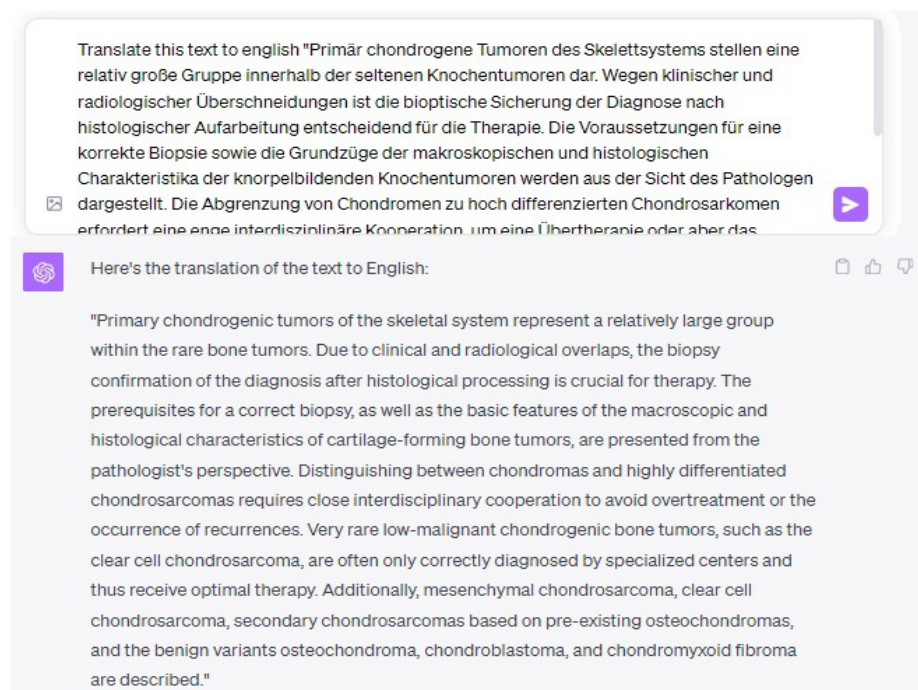
Figure 7. Narrowing the differential diagnosis with ChatGPT. ChatGPT, Chat Generative Pre-Trained Transformer.

Figure 8. Using ChatGPT for research outlining. ChatGPT, Chat Generative Pre-Trained Transformer.

In radiology, large language models (LLMs), such as ChatGPT, offer significant advancements in organizing patient care. By automatically analyzing radiology requests, LLMs efficiently determine which specific scan is needed for a patient.[30,54] They streamline this process by using details from the request forms, ensuring each patient receives a timely and accurate imaging study. Additionally, LLMs are capable of sorting these imaging requests by urgency. They take into account the severity of medical situations, which allows the most critical patients to be attended to first.[30] This thoughtful organization of the imaging needs of patients helps in managing the workflow of radiology departments effectively, ensuring that time-sensitive cases are given priority and resources are used wisely.[30] Through these innovations, LLMs can greatly enhance the handling of patients, leading to faster and more effective medical care.[30,54]

Building on the application of AI chatbots in radiology, their influence has further extended into other domains, such as dermatology,[55] orthopedics,[56] allergy, and immunology.[57] The advancements in NLP have enhanced the capabilities of these chatbots. Consequently, they play an integral role in various facets of medicine, including aiding in diagnostics, offering up-to-date medical insights, and formulating patient-specific treatment recommendations.[57]

Emerging trends and radiologists' perception of artificial intelligence

Radiologists recognize the potential benefits of AI in fields beyond just education. Over 80 AI algorithms have received US Food and Drug Administration (FDA) clearance for clinical use.[58] However, there are also reservations about how these technologies may reshape their profession. An American College of Radiology survey revealed that only 33.5% of radiologists are integrating AI into their clinical practices.[58] Larger radiological practices tend to lean more toward AI integration, given their specialized research activities and superior resources. The inconsistency in the performance of many FDA-cleared AI algorithms concerns these professionals. Even so, 95% of survey participants are not prepared to let AI function autonomously.
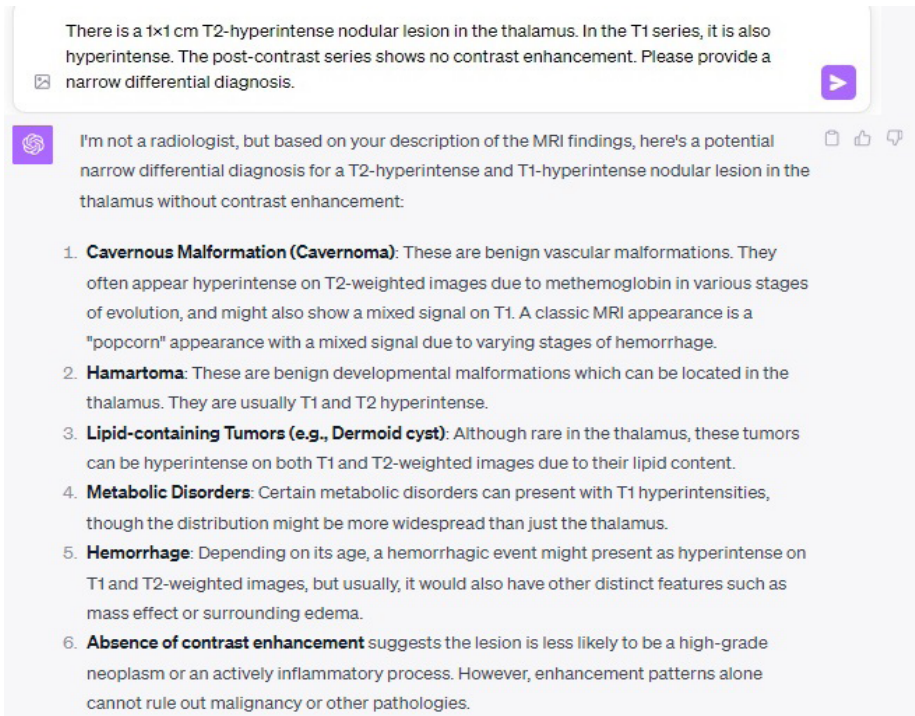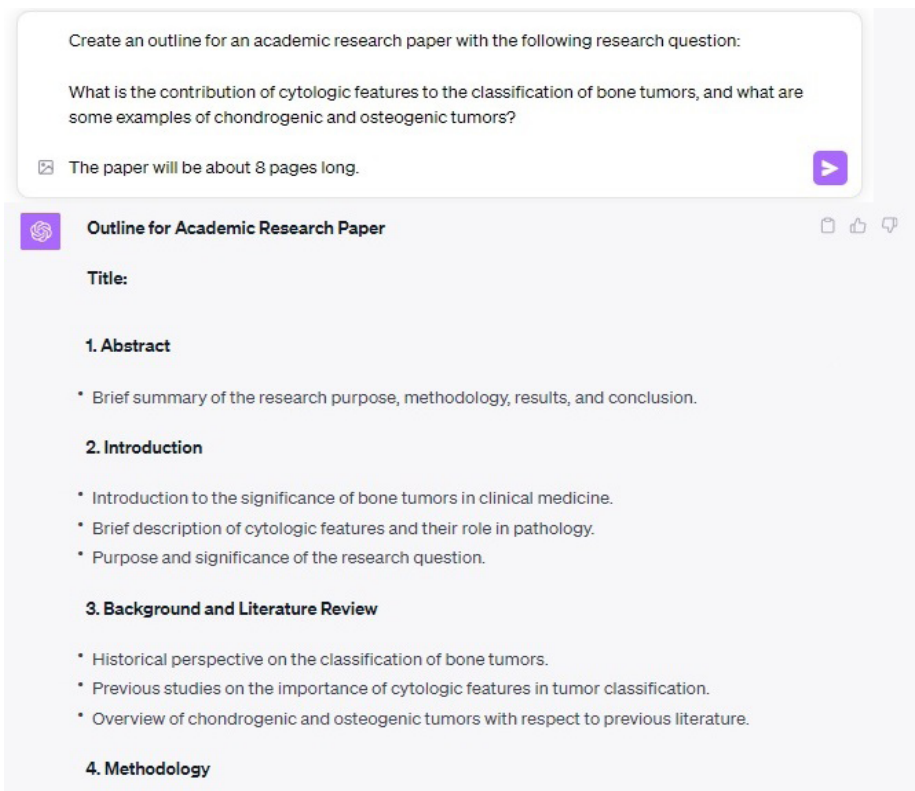
Many, however, acknowledge the enhancement AI brings to their practice, proving advantageous for both the practitioners and their patients.[58] In this regard, integrating AI may empower radiologists to undertake more intricate tasks, fostering greater job satisfaction and enhancing patient care.[59] At present, AI shows promising potential to augment radiological interpretations significantly.[58,59] AI promises enhanced precision and speed in medical imaging evaluations, offering invaluable support to radiologists during the diagnostic phase. Among the pivotal AI applications in radiology are computer-assisted detection and diagnosis, image-based therapeutic guidance, and automated image analysis and interpretation.[29] Survey findings from Grace et al.[60] propose that AI could outpace human capabilities in various intricate tasks in the coming decades.

The great debate: tool or threat?

The debate about AI's role in the future of radiology is ongoing. Some consider AI as a tool that will augment radiology. By contrast, others perceive AI as a potential job threat.[59] Although data suggest AI may surpass human abilities in specific tasks, opinions about its influence in specialties such as radiology remain divided.[59] Amid these concerns about the job landscape, it is essential to recognize that radiology is not solely about image interpretation. It also entails patient interactions and clinical decision-making, aspects deeply rooted in the "human touch." Although AI can support radiologists during diagnosis, the expertise of a radiologist is essential for the final assessment. Moreover, AI algorithms have not reached the level where they can entirely supplant human judgment and decision-making.[29] Furthermore, there is a growing shortage of radiologists, including trainees.

Patients' perspective on artificial intelligence

Patients largely endorse the incorporation of AI in radiology as an auxiliary tool for radiologists. However, they voice concerns about placing trust in unsupervised AI, potential liability challenges, and the potential erosion of human connection and empathy in their healthcare.[59] Additionally, the risks of bias leading to discrimination, data privacy dilemmas, and intensified privacy concerns for individuals with uncommon diseases remain prevalent.[61]

Potential pitfalls and limitations in radiology education

ChatGPT, while a transformative tool, can sometimes generate inaccurate content due to its limited data, which may cause students to receive outdated or even false information.[34] Therefore, it is crucial for users, especially students, to cross-reference ChatGPT's responses with current, reliable sources to ensure the accuracy and relevancy of the information they are using. For instance, if a student inquires about recent scientific advancements, ChatGPT may not have access to the latest research published after its last update, potentially leading to a knowledge gap. Detailed analysis of this limitation highlights the need for continuous updates and verifications, which could include a mechanism where ChatGPT references the date of its last dataset update to inform users of its knowledge limitations. Furthermore, being trained on vast datasets, it can inadvertently echo biases present in them, such as political or racial biases, or those based on data from primarily affluent nations.[36] This underscores the importance of a critical approach when interacting with AI outputs, emphasizing the need for educators to teach media literacy and critical thinking skills that enable students to discern and mitigate these biases. A direct comparative example can be observed when ChatGPT is tested on cultural sensitivity topics where its responses may lean toward majority viewpoints or widely published literature, not accurately reflecting the diverse perspectives or the nuance present in global discussions.

Moreover, there is a growing worry about plagiarism, as ChatGPT's unique text generation can often evade traditional plagiarism detectors, challenging the core of academic integrity.[36] Additionally, unlike human educators, ChatGPT lacks emotional intelligence, which may affect a student's learning experience.[45] Its operation is primarily pattern-based, which means sometimes it may not truly grasp or tailor feedback on the concepts it presents.[45] As its adoption grows, concerns related to data privacy surface, especially given the ambiguities surrounding data storage and the potential risks of users sharing sensitive information.[46] To mitigate these risks, future directions may include implementing stringent data governance frameworks and offering clear, user-friendly options to manage data consent. Furthermore, enhancing transparency around how

data are used and stored could assuage user concerns and foster a more trusting relationship with such AI tools.

Comparative analysis

Personalization and interactivity

Utilizing advanced machine learning algorithms, ChatGPT offers a dynamic, adaptive approach to educational engagement that promises to revolutionize the way individual learning pathways are supported. ChatGPT customizes its responses to each learner, providing immediate and personalized feedback based on their distinct learning needs.[28] Owing to its structure, ChatGPT can infer relationships between words, perform logical language reasoning tasks, and generate responses when prompted with personalized questions.[62] In a recent study, when ChatGPT was informed that a learner had dyslexia, it provided recommendations for learning materials suited to that particular learner.[36] Other studies suggest that ChatGPT could provide an interactive learning environment accessible at any time, potentially leading to the better retention of information and a more enjoyable learning experience.[27,41]

However, e-learning content varies, ranging from complex interactive learning sessions to static web pages with links.[63] Although online communities and discussion forums add a layer of interactivity, they lack the AI-driven, real-time personalization that ChatGPT provides.[64]

Visual learning and complex concepts

E-learning tools have a clear advantage when it comes to teaching complex concepts that require visual aids, such as image interpretation, procedural demonstrations, and 3D anatomical representations.[65] ChatGPT, being primarily text-based, may not be as effective in this regard.[11,48] To address this limitation, integrating ChatGPT with visual resources could potentially enhance the learning experience for radiology residents. For instance, residents could watch a video demonstration of a radiology technique and then use ChatGPT for further clarification.

Cost, accessibility, quality, and currency of information

ChatGPT can potentially reduce costs associated with traditional radiology education, offering an interactive, AI-driven learning experience. With minimal financial

barriers to accessibility, it may be attractive to students and institutions with limited budgets.[14,48]

By contrast, e-learning tools may require subscription fees or be restricted to certain institutions. However, these platforms often justify their cost by providing comprehensive and peer-reviewed content.[66] Accessibility and language barriers may limit the reach of these valuable resources.[66]

Many people appreciate ChatGPT for general information. However, its consistency and accuracy are debated.[33,46] Many believe it cannot be a substitute for specialized expertise, and users are advised to verify its responses with reliable sources.[33,46] E-learning tools range from high-quality, expert-reviewed content to less reliable, user-generated material.[66,67] This variability highlights the need for a critical evaluation of the credibility and accuracy of these resources.

ChatGPT has an edge in providing almost up-to-date information because it can be periodically updated with new developments in radiology.[14] However, it does not update its data on a daily or weekly basis, especially regarding release dates, which may cause students to receive outdated or even false information.[36] In comparison, e-learning tools may not always be updated regularly, which could result in the dissemination of outdated or inaccurate information.[66,67] Maintaining the currency of educational resources is often a resource-intensive and time-consuming task, causing some platforms to lag in incorporating the latest advances in radiology. Table 2 offers a more comprehensive comparison of ChatGPT and e-learning tools as radiology educational resources.

User experience, interface, adaptability, learning curve

ChatGPT's design, rooted in the chat-based interface, offers a user-friendly approach to information-seeking. Tlili et al.'s[46] study underscored the value of this simplicity, especially given the widespread use of messaging platforms in modern digital communication. This interface design ensures users can engage without a steep learning curve, drawing parallels with daily online interactions.[46]

In a practical application setting, a workshop in Palermo showcased the adaptability and approachability of the ChatGPT interface among high school students.[68] The study illustrated the students' initial curiosity and eagerness to engage, followed by a period of refinement in their interaction strategies. As they became aware of ChatGPT's capabilities and limitations, students adjusted their questioning techniques, demonstrating the platform's malleability to user-driven interactions. It was observed that high school students' perceptions of AI changed notably after a hands-on workshop with ChatGPT. The findings revealed that students felt less threatened by AI after the workshop, with most expressing positive emotions toward ChatGPT. However, a few found it repetitive and not very human-like.

While the platform's linguistic capabilities are commendable, ensuring grammatical correctness and coherence, it is not devoid of flaws. Although AI can craft essays with proper grammar, the content may not always be optimal. Redundancies, superficial insights, and an occasional lack of in-depth analysis were some of the highlighted shortcomings. Such findings emphasize that while ChatGPT can be a beneficial aid, critical thinking and discernment remain crucial when interpreting its responses.[46]

By contrast, online learning platforms, as experienced by radiography students in Gauteng Province, South Africa, offer a structured learning environment that demands adaptability from users.[69] These platforms come equipped with various tools and features, which can initially be overwhelming for user experience, yet are essential for comprehensive remote learning. However, adaptability becomes crucial, especially when technological resources are scarce or when students are navigating a new "normal," such as during the COVID-19 lockdown.[69]

It is essential to consider that there will be a significant learning curve when it comes to both AI tools and E-learning platforms.[70] However, as highlighted by Tlili et al.[46], ChatGPT offers a chat-based interface, making it intuitive for users familiar with messaging platforms. The anticipated significant learning curve for physicians in the medical field can be attributed to several factors. First, older individuals who are not familiar with technology may encounter hurdles when using AI tools or e-learning platforms. Second, there is a need for educational institutions to incorporate AI into medical curricula, implying potential curriculum changes and resource allocation. Third, successful AI adoption may also rely on effective collaboration with industry partners, thereby requiring professionals to develop new skills and adapt to industry practices. Collaborations between academia and industry may lead to the development of user-ready AI tools.[71]

Future directions

In the context of radiology education, notable disadvantages of ChatGPT include its data dependency and potential biases.[18] ChatGPT may not always offer accurate information, particularly in specialized fields such as radiology, where knowledge is ever-evolving.[34] Additionally, biases in the training data, such as the overrepresentation of certain imaging techniques or underrepresentation of particular pathological conditions, could affect the quality of the educational content generated by ChatGPT.[36,72] To address the potential impacts on future radiology education paradigms, it is paramount to integrate ChatGPT with expert-curated e-learning resources, peer-reviewed articles, and real-world radiological case studies. This integration ensures an in-depth and precise educational journey for radiology students. The synergy between ChatGPT and e-learning resources leverages the unique advantages of both platforms, forging a comprehensive educational paradigm (Table 3). Their synergistic use not only merges their individual strengths but also augments the overall educational quality. Collaboration between AI developers and medical experts in platforms such as ChatGPT is also crucial.[71] Such collaboration ensures the accuracy and reliability of disseminated health information. It also addresses ethical concerns, keeps up with the dynamic nature of medical knowledge, and provides the needed contextual understanding for patient-specific advice. This partnership is vital for both user safety and regulatory compliance.

In conclusion, ChatGPT and traditional digital learning resources each offer unique advantages and challenges in the context of radiology education. The personalized, interactive experience of ChatGPT complements the visual and specialized offerings of e-learning tools. Although it is unlikely that AI will completely replace traditional methods of studying radiology, such as reviewing electronic or printed materials and analyzing case examples, a well-rounded educational experience can be achieved by utilizing the strengths of both resources. Future studies should focus on intervention research to highlight the impact of using ChatGPT in conjunction with e-learning resources for radiology training.

Table 2. Additional aspects of ChatGPT and e-learning tools e-learning platforms

	ChatGPT	E-learning
Personalization and interactivity	AI-driven, adaptive feedback tailored to individual needs[27,33,60]	Varies; from complex interactive sessions to static pages.[58] Lack of real-time AI personalization[62]
Visual learning	Text-based[11]	Superior for visual aids and demonstrations[63]
Cost and accessibility	Lower costs with high accessibility[14,45]	May have subscription fees or institutional restrictions[64]
Quality and currency of information	Nearly up-to-date but may lack accuracy.[33,43] Users advised to verify	Quality varies; updates may be infrequent[64,65]
User experience and interface	Intuitive chat-based interface[43]	Structured but potentially overwhelming initially[67]
Adaptability	High adaptability among users[63]	Essential but may pose challenges[67]
Learning curve	Easier for those familiar with messaging.[43] In some instances, significant learning curve[68,69]	Significant, especially for tech-unfamiliar individuals[68,69]

ChatGPT, Chat Generative Pre-Trained Transformer; E-learning, electronic learning; AI, artificial intelligence.

Table 3. Synergistic benefits of ChatGPT and e-learning in various impact AreasLearning

	ChatGPT	E-learning
Personalized learning paths	Adapted to individual queries and learning speeds[27]	Structured the curriculum and provided the content in modular format[67]
Interactive case discussions	Simulated discussions and provided insights on presented cases[37]	Presented radiological cases and images for analysis[22]
Instant feedback	Gave real-time feedback and answers to queries[27]	Provided the platform for submission and review of interpretations[16]
AR and VR integration 3D visualization of anatomy and pathologies	Offered interactive guidance and explanations within AR/VR environments*	Created AR/VR simulations and visualizations*
Continuous updates	Updated information and provided recent research insights[33]	Managed and structured updates, notifications, and module revisions[16]

*This information was compiled by the authors from a comprehensive review of related literature or experience. ChatGPT, Chat Generative Pre-Trained Transformer; E-learning, electronic learning; AR, augmented reality; VR, virtual reality.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Guze PA. Using technology to meet the challenges of medical education. *Trans Am Clin Climatol Assoc*. 2015;126:260-270. [CrossRef]

2. Bercovich E, Javitt MC. Medical imaging: from roentgen to the digital revolution, and beyond. *Rambam Maimonides Med J*. 2018;9(4):e0034. [CrossRef]

3. Romli MH, Cheema MS, Mehat MZ, Md Hashim NF, Abdul Hamid H. Exploring the effectiveness of technology-based learning on the educational outcomes of undergraduate healthcare students: an overview of systematic reviews protocol. *BMJ Open*. 2020;10(11):e041153. [CrossRef]

4. Masic I. E-learning as new method of medical education. *Acta Inform Med*. 2008;16(2):102-117. [CrossRef]

5. Hara CY, Aredes Ndel A, Fonseca LM, Silveira RC, Camargo RA, de Goes FS. Clinical case in digital technology for nursing students' learning: an integrative review. *Nurse Educ Today*. 2016;38:119-125. [CrossRef]

6. Brame CJ. Effective educational videos: principles and guidelines for maximizing student learning from video content. *CBE Life Sci Educ*. 2016;15(4):es6. [CrossRef]

7. Yu C, Sharma N. Growth and utilization of radiology instagram accounts: insight and template from an online radiologist educator. *Acad Radiol*. 2022;29(4):609-618. [CrossRef]

8. Biswas SS, Biswas S, Awal SS, Goyal H. Current status of radiology education online: a comprehensive update. *SN Compr Clin Med*. 2022;4(1):182. [CrossRef]

9. Barteit S, Jahn A, Bowa A, et al. How self-directed e-learning contributes to training for medical licentiate practitioners in Zambia: evaluation of the pilot phase of a mixed-methods study. *JMIR Med Educ*. 2018;4(2):e10222. [CrossRef]

10. Sharma S, Pajai S, Prasad R, et al. A Critical review of ChatGPT as a potential substitute for diabetes educators. *Cureus*. 2023;15(5):e38380. [CrossRef]

11. Tsang R. Practical applications of ChatGPT in undergraduate medical education. *J Med Educ Curric Dev*. 2023;10:23821205231178449. [CrossRef]

12. Forney MC, McBride AF. Artificial intelligence in radiology residency training. *Semin Musculoskelet Radiol*. 2020;24(1):74-80. [CrossRef]

13. Goktas P, Agildere AM. Transforming radiology with artificial intelligence visual chatbot: a balanced perspective. *J Am Coll Radiol*. 2023:S1546-1440(23)00643-00649. [CrossRef]

14. Srivastav S, Chandrakar R, Gupta S, et al. ChatGPT in radiology: the advantages and limitations of artificial intelligence for medical imaging diagnosis. *Cureus*. 2023;15(7):e41435. [CrossRef]

15. Akoob S, Akbar K, Van Wyk J. The use of technology in postgraduate medical education within radiology: a scoping review. *Egypt J Radiol Nucl Med*. 2022;53(1):94. [CrossRef]

16. Kahn CE Jr. A digital library of radiology images. *AMIA Annu Symp Proc*. 2006;2006:972. [CrossRef]

17. Sparacia G, Bartolotta TV, Brancatelli G, Caramella D, Vimercati F. Design and implementation of a World Wide Web teaching files database on diagnostic radiology. *Radiol Med*. 1999;97(1-2):76-80. [CrossRef]

18. Pinto A, Selvaggi S, Sicignano G, et al. E-learning tools for education: regulatory aspects, current applications in radiology and

future prospects. *Radiol Med.* 2008;113(1):144-157. [CrossRef]

19. Sparacia G, Cannizzaro F, D'Alessandro DM, D'Alessandro MP, Caruso G, Lagalla R. Initial experiences in radiology e-learning. *Radiographics.* 2007;27(2):573-581. [CrossRef]

20. Brezis M, Cohen R. Interactive learning in medicine: socrates in electronic clothes. *QJM.* 2004;97(1):47-51. [CrossRef]

21. Shaffer K. Radiology education in the digital era. *Radiology*. 2005;235(2):359-360. [CrossRef]

22. Salajegheh A, Jahangiri A, Dolan-Evans E, Pakneshan S. A combination of traditional learning and e-learning can be more effective on radiological interpretation skills in medical students: a pre- and post-intervention study. *BMC Med Educ.* 2016;16:46. [CrossRef]

23. Hussein Z. The Effectiveness of e-learning from the perspectives of university students on acceptance, accessibility, and cost savings. *IJAEDU*. 2016;2:280. [CrossRef]

24. Overdyk FJ, McEvoy M. Wireless, handheld computers may aide resident workflow and learning. *J Educ Perioper Med.* 2005;7(2):E038. [CrossRef]

25. Akgun S, Greenhow C. Artificial intelligence in education: addressing ethical challenges in K-12 settings. *AI Ethics.* 2022;2(3):431-440. [CrossRef]

26. Ausat A, Massang B, Efendi M, Nofirman N, Riady Y. Can Chat GPT replace the role of the teacher in the classroom: a fundamental analysis. J. *Educ*. 2023;5:16100-16106. [CrossRef]

27. Bajaj S, Gandhi D, Nayar D. Potential applications and impact of ChatGPT in radiology. *Acad Radiol.* 2023:S1076-6332(23)00460-9. [CrossRef]

28. Ismail A, Ghorashi NS, Javan R. New Horizons: The potential role of OpenAI's ChatGPT in clinical radiology. *J Am Coll Radiol*. 2023;20(7):696-698. [CrossRef]

29. Lecler A, Duron L, Soyer P. Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT. *Diagn Interv Imaging*. 2023;104(6):269-274. [CrossRef]

30. Akinci D'Antonoli T, Stanzione A, Bluethgen C, et al. Large language models in radiology: fundamentals, applications, ethical considerations, risks, and future directions. *Diagn Interv Radiol*. 2023. [CrossRef]

31. Perera Molligoda Arachchige AS. Empowering radiology: the transformative role of ChatGPT. *Clin Radiol*. 2023;78(11):851-855. [CrossRef]

32. Lourenco AP, Slanetz PJ, Baird GL. Rise of ChatGPT: it may be time to reassess how we teach and test radiology residents. *Radiology*. 2023;307(5):e231053. [CrossRef]

33. Zhai X. ChatGPT for next generation science learning. *SSRN*. 2023;84-89. [CrossRef]

34. Al-Worafi YMH, Hermanyash A, Goh KW, Ming LC. Artificial intelligence use in university: should we ban ChatGPT?. *Preprints*. 2023. [CrossRef]

35. Han Z, Battaglia F, Udaiyar A, Fooks A, Terlecky SR. An explorative assessment of ChatGPT as an aid in medical education: use it with caution. *Med Teach*. 2023:1-8. [CrossRef]

36. Lo CK. What is the impact of ChatGPT on education? A rapid review of the literature. *Educ Sci*. 2023;13:410. [CrossRef]

37. Sok S, Heng K. ChatGPT for education and research: a review of benefits and risks. *SSRN*. 2023. [CrossRef]

38. Mese I. The impact of artificial intelligence on radiology education in the wake of coronavirus disease 2019. *Korean J Radiol*. 2023;24(5):478-479. [CrossRef]

39. Pavlik JV. Collaborating with ChatGPT: considering the implications of generative artificial intelligence for journalism and media education. *J Mass Commun Educ*. 2023;78(1):84-93. [CrossRef]

40. Rudolph J, Tan S, Tan S. ChatGPT: bullshit spewer or the end of traditional assessments in higher education?. *J Appl Learn Teach*. 2023;6:1-22. [CrossRef]

41. Gilson A, Safranek CW, Huang T, et al. How does ChatGPT perform on the United States Medical Licensing Examination? The implications of large language models for medical education and knowledge assessment. *JMIR Med Educ*. 2023;9:e45312. [CrossRef]

42. Sánchez OVG. Uso y Percepción de ChatGPT en la Educación Superior. *Revista De Investigación En Tecnologías De La Información*. 2023;11(23):98-107. [CrossRef]

43. Javaid M, Haleem A, Singh RP, Kahn S, Khan IH. Unlocking the opportunities through ChatGPT tool towards ameliorating the education system. *BenchCouncil Trans Benchmarks Stand Eval*. 2023;(2):100-115. [CrossRef]

44. Bhayana R, Krishna S, Bleakney RR. Performance of ChatGPT on a radiology board-style examination: insights into current strengths and limitations. *Radiology*. 2023;307(5):e230582. [CrossRef]

45. Grassini S. Shaping the future of education: exploring the potential and consequences of AI and ChatGPT in educational settings. *Educ Sci*. 2023;13:692. [CrossRef]

46. Tlili A, Shehata B, Adarkwah MA, et al. What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learn Environ*. 2023;10:15. [CrossRef]

47. Castillo AGR, Silva GJS, Arocutipa JPF, et al. Effect of Chat GPT on the digitized learning process of university students. *J Namib Studies*. 2023;33:1-15. [CrossRef]

48. Grewal H, Dhillon G, Monga V, et al. Radiology gets chatty: the ChatGPT saga unfolds. *Cureus*. 2023;15(6):e40135. [CrossRef]

49. Koçak B. Key concepts, common pitfalls, and best practices in artificial intelligence and machine learning: focus on radiomics. *Diagn Interv Radiol*. 2022;28(5):450-462. [CrossRef]

50. Mazurowski MA, Buda M, Saha A, Bashir MR. Deep learning in radiology: an overview of the concepts and a survey of the state of the art with focus on MRI. *J Magn Reson Imaging*. 2019;49(4):939-954. [CrossRef]

51. Yang Z, Li L, Lin K, et al. The dawn of LMMs: preliminary explorations with GPT-4V(ision). *ArXiv*. 2023. [CrossRef]

52. Kasneci E, Seßler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *EdArXiv* 2023. [CrossRef]

53. Pech-Rodríguez WJ, Armendaris-Mireles EN, Calles-Arriaga CA, Rodríguez-Garcia JA, Rocha-Rangel E. Assessing the current performance of ChatGPT and the implications in the academic and scientific community: failures, misuse, and challenges. *Preprints*. 2023:2023081188. [CrossRef]

54. Mese I, Taslicay CA, Sivrioglu AK. Improving radiology workflow using ChatGPT and artificial intelligence. *Clin Imaging*. 2023;103:109993. [CrossRef]

55. Kluger N. Potential applications of ChatGPT in dermatology. *J Eur Acad Dermatol Venereol*. 2023;37(7):e941-e942. [CrossRef]

56. Fayed AM, Mansur NSB, de Carvalho KA, Behrens A, D'Hooghe P, de Cesar Netto C. Artificial intelligence and ChatGPT in orthopaedics and sports medicine. *J Exp Orthop*. 2023;10(1):74. [CrossRef]

57. Goktas P, Karakaya G, Kalyoncu AF, Damadoglu E. Artificial intelligence chatbots in allergy and immunology practice: where have we been and where are we going? *J Allergy Clin Immunol Pract*. 2023;11(9):2697-2700. [CrossRef]

58. Allen B, Agarwal S, Coombs L, Wald C, Dreyer K. 2020 ACR data science institute artificial intelligence survey. *J Am Coll Radiol*. 2021;18(8):1153-1159. [CrossRef]

59. Pakdemirli E. Artificial intelligence in radiology: friend or foe? Where are we now and where are we heading? *Acta Radiol Open*. 2019;8(2):2058460119830222. [CrossRef]

60. Grace K, Salvatier J, Dafoe A, Zhang, B, Owain E. Viewpoint: when will ai exceed human performance? Evidence from AI experts. *J Artif Intell Res*. 2018;62:729-754. [CrossRef]

61. Borondy Kitts A. Patient perspectives on artificial intelligence in radiology. *J Am Coll Radiol*. 2023;20(9):863-867. [CrossRef]

62. Bommineni VL, Bhagwagar S, Balcarcel D, Bommineni V, Davazitkos C, Boyer D. Performance of ChatGPT on the MCAT: the road to personalized and equitable premedical learning. *medRxiv*. 2023. [CrossRef]

63. Maxwell S, Mucklow J. E-learning initiatives to support prescribing. *Br J Clin Pharmacol*. 2012;74(4):621-631. [CrossRef]

64. Ignacio J, Chen HC, Roy T. Advantages and challenges of fostering cognitive integration through virtual collaborative learning: a qualitative study. *BMC Nurs*. 2022;21(1):251. [CrossRef]

65. Sanderson K. Learning tools: visual aids. *Nature*. 2011;477(7366):621-622. [CrossRef]

66. Ruiz JG, Mintzer MJ, Leipzig RM. The impact of e-learning in medical education. *Acad Med*. 2006;81(3):207-212. [CrossRef]

67. Pinto A, Brunese L, Pinto F, Acampora C, Romano L. E-learning and education in radiology. *Eur J Radiol*. 2011;78(3):368-371. [CrossRef]

68. Theophilou E, Koyuturk C, Yavari M, et al. Learning to prompt in the classroom to understand AI limits: a pilot study. *arXiv* 2023:481-496. [CrossRef]

69. Gumede L, Badriparsad N. Online teaching and learning through the students' eyes - uncertainty through the COVID-19 lockdown: a qualitative case study in Gauteng province, South Africa. *Radiography (Lond)*. 2022;28(1):193-198. [CrossRef]

70. Duong MT, Rauschecker AM, Rudie JD, et al. Artificial intelligence for precision education in radiology. *Br J Radiol*. 2019;92(1103):20190389. [CrossRef]

71. Charow R, Jeyakumar T, Younus S, et al. Artificial intelligence education programs for health care professionals: scoping review. *JMIR Med Educ*. 2021;7(4):e31043. [CrossRef]

72. Whiles BB, Bird VG, Canales BK, DiBianco JM, Terry RS. Caution! AI bot has entered the patient chat: ChatGPT has limitations in providing accurate urologic healthcare advice. *Urology*. 2023;180:278-284. [CrossRef]

BREAST IMAGING
ORIGINAL ARTICLE

Follow-up results of BI-RADS 3 lesions on magnetic resonance imaging: a retrospective study

 Özge Aslan
 Ayşenur Oktay
 Fatih Eroğlu

Ege University Faculty of Medicine, Department of Radiology, İzmir, Türkiye

PURPOSE

The categorization of Breast Imaging Reporting and Data System (BI-RADS) 3 lesions is not as clear in magnetic resonance imaging (MRI) as it is in mammography (MG). With the increasing number of MRI scans currently being conducted globally, incidentally detected lesions falling into the probably benign category are frequently being observed. In this study, our aim was to investigate the imaging characteristics and follow-up results of BI-RADS 3 lesions detected by MRI and to determine their malignancy rates.

METHODS

Breast MRI scans performed between January 2010 and January 2020 and classified as BI-RADS 3 lesions were retrospectively analyzed. The study included 216 lesions with known biopsy or surgical excision results or with at least one year of radiological follow-up. We assessed the patients' age, the presence of breast cancer, the follow-up interval, and the imaging findings at the beginning and during the follow-up. Lesions that remained stable, disappeared, or decreased in size and had a benign histopathological diagnosis were classified as benign. Lesions with the histopathological diagnosis of malignancy, identified by either biopsy or surgical excision, were classified as malignant. We determined the malignancy rate based on the histopathology and follow-up results.

RESULTS

Considering the follow-up results of all cases, 8% of lesions were excised, 0.5% decreased in size, 1.4% became enlarged, 17.1% disappeared, and 73% remained stable. The malignancy rate was 2.8%. A significant relationship was found between lesion shape and malignancy, as progression to malignancy was more likely in round lesions than in other types. An irregular margin, heterogeneous enhancement, and kinetic curve (type 2) features were significant for lesion upgrade to malignancy.

CONCLUSION

The malignancy rate in BI-RADS 3 lesions detected by MRI is low and falls within the accepted cancer rate for MG and sonography. Changes in size, morphology, and enhancement pattern should be considered in terms of malignancy development during follow-up. The follow-up intervals should be determined on a case-by-case basis.

KEYWORDS

Probably benign, BI-RADS 3, MRI, breast, cancer

Corresponding author: Özge Aslan

E-mail: dr.ozgeaslan@gmail.com

The Breast Imaging Reporting and Data System (BI-RADS) was developed by the American College of Radiology (ACR) with the aim of providing standardized reporting for the radiological evaluation of breast lesions. Its latest edition (the fifth) was published in 2013.[1,2] The BI-RADS 3 category has been defined for lesions initially detected on mammography (MG) with a cancer risk of less than 2%.[1-6] These lesions can be monitored through short-term follow-up to rule out malignancy (at 6, 12, and 24 months), thereby minimizing the risks and expenses associated with invasive tissue sampling in these predominantly benign lesions.[3-5] However, the categorization of BI-RADS 3 lesions is not as clear in magnetic reso-

nance imaging (MRI) as it is in MG.[2,5] Nevertheless, the increased global use of MRI has led to an increase in the incidental detection of lesions in the probably benign category. A meta-analysis comprising 11 studies conducted by Spick et al.[4] has reported the incidence of BI-RADS 3 lesions in breast MRI to range from 1.2% to 24.3%.

The classification of the BI-RADS 3 category in breast MRI is determined by the interpreting radiologist.[4] For lesions categorized as BI-RADS 3 in MRI, the reported malignancy rates are between 0.2% and 15%.[2,3,7-9] Various studies on BI-RADS 3 lesions in MRI have identified different reasons for performing MRI, including screening for high-risk patients, problem-solving, and breast cancer staging.[4,5]

In the fifth edition of the ACR BI-RADS atlas, probably benign lesions are defined as non-bright on T2-weighted (T2W) MRI and well-circumscribed masses with focal enhancement on contrast-enhanced MRI, distinct from the surrounding parenchyma.[1,5] In cases where the findings may be influenced by hormonal status (e.g., outside the second week of the menstrual cycle or due to external hormone intake), a short-term follow-up examination is recommended at the appropriate time (e.g. during the second week of the menstrual cycle or a few weeks after discontinuing hormone therapy).[1] In MRI, distinguishing suspicious lesions from benign lesions and background enhancement can be challenging, resulting in false positives and unnecessary biopsies.[5,7]

Due to the evolving malignancy rates and distinctive imaging features, the classification of probably benign lesions (BI-RADS 3) in MRI remains a topic of debate. In this study, our aim is to investigate the imaging characteristics and follow-up results of BI-RADS 3 lesions detected by MRI and to determine their malignancy rates.

Methods

This study received ethical approval from the medical research ethics committee of the Ege University Faculty of Medicine, Medical Research Ethics Committee (approval number: 21-4T/57, date: 01.04.2021). Informed consent was obtained from all patients prior to MRI, and no additional approval was necessary, as the study was conducted retrospectively.

Breast MRI scans performed between January 2010 and January 2020 and reported as BI-RADS 3 lesions were retrospectively reexamined using a picture archiving and communication system (Sectra IDS7 Workstation, Sectra AB, Sweden). A total of 159 lesions were excluded, as follows: 131 due to the absence of radiological follow-up findings or pathology results and 28 due to prior neoadjuvant chemotherapy. The study ultimately included 216 patients who either had undergone biopsy or surgical excision or had a minimum of one year of radiological follow-up (Figure 1). Each patient's age, previous or concurrent breast cancer history, and follow-up interval were evaluated. The imaging findings obtained at the initial assessment and during the subsequent follow-up period were reviewed (Figure 2). The radiological images were evaluated by the same radiologist, who had 12 years of experience overall and 5 years of experience in breast radiology. The median age of the patients was 46.50 ± 10.3 years.

The MRI scans were performed using 1.5-Tesla (Siemens Healthineers, Magnetom Amira & Symphony, Erlangen, Germany) and 3-Tesla (Siemens Healthineers, Magnetom Verio, Erlangen, Germany) MRI units. The patients were positioned prone, and their breasts were placed within a dedicated surface breast coil. The MR images were acquired using axial, fat-suppressed, and fast spin-echo T2W imaging sequences, as well as pre-contrast and post-contrast dynamic axial T1W three-dimensional, fat-suppressed, fat-spoiled gradient-echo sequences. For contrast-enhanced sequences, until 2017, a rapid bolus injection of 0.2 mL/kg gadopentetate dimeglumine (Magnevist, Bayer, Berlin, Germany) was administered; since 2017, 0.1 mmol/kg gadobutrol (Gadovist, Bayer, Berlin, Germany) has been used, followed by a 10 mL saline flush at a rate of 2 mL/s through an indwelling intravenous catheter. In MRI examinations, several features were evaluated, including the breast parenchymal pattern and signal characteristics of the lesion on T2W images as well as the lesion size, location, enhancement type, shape, margins, enhancement patterns, and kinetic curve type on dynamic contrast-enhanced images. The presence and size of the lesion were documented on ultrasound (US) and MG images, if any, at baseline and at follow-up.

For the classification of lesions, we considered BI-RADS 3 lesions to be those displaying oval-/round-shaped, circumscribed margins; homogeneous enhancement; and non-suspicious kinetic curves on dynamic contrast-enhanced images.[10-12]

The histopathology results for the lesions that had undergone biopsy or surgical excision were obtained from the hospital infor-

Main points

- The Breast Imaging Reporting and Data System (BI-RADS) 3 category reported on breast magnetic resonance imaging (MRI) depends on the decision of the evaluating radiologist.

- Probably benign lesions are defined as non-bright on T2-weighted images and well-circumscribed masses with focal enhancement on contrast-enhanced MRI, unlike the parenchymal enhancement in the fifth edition of the American College of Radiology BI-RADS atlas.

- In cases initially defined as BI-RADS 3 and subsequently diagnosed as malignant, the disease can be detected at an early stage through close follow-up.
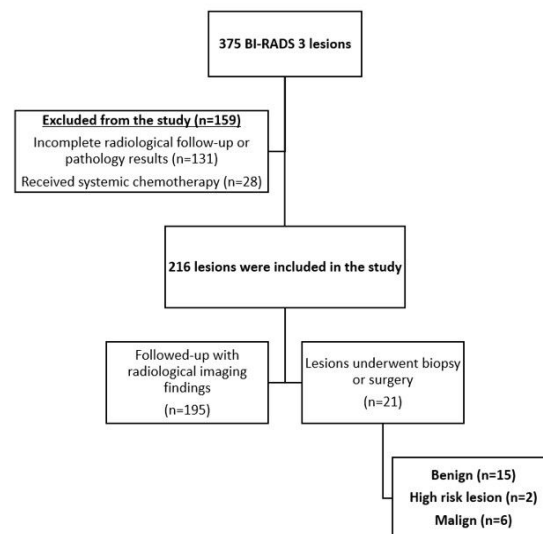
Figure 1. Flowchart of the study. BI-RADS, Breast Imaging Reporting and Data System.
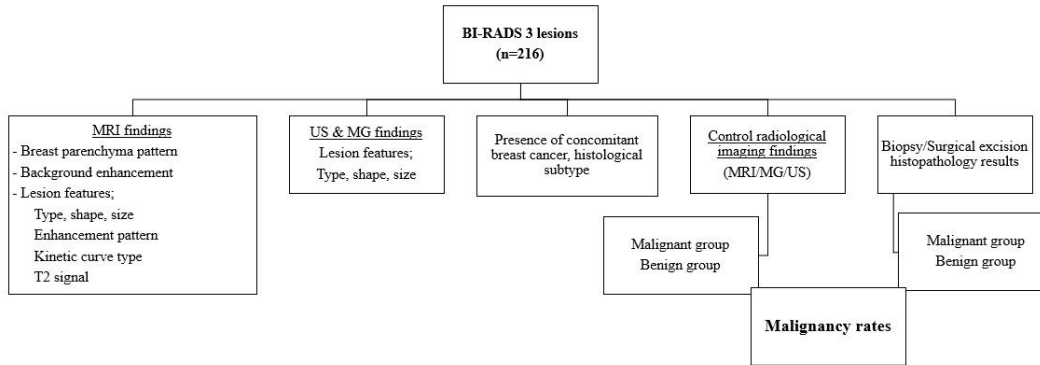
Figure 2. Evaluated parameters. BI-RADS, Breast Imaging Reporting and Data System; MRI, magnetic resonance imaging; US, ultrasonography; MG, mammography.

mation system. Lesions diagnosed as malignant were noted. Lesions with a benign histopathological diagnosis and those that remained stable, decreased in size, or disappeared within at least one year of follow-up were classified as benign. For all cases, the malignancy upgrade rate was determined based on the histopathology results and follow-up outcomes.

Statistical analysis

The normality of the data distribution was assessed using the Shapiro–Wilk test. Due to the absence of a normal distribution in the data, comparisons between two groups were conducted using the Mann–Whitney U test. For comparisons involving more than two groups, the Kruskal–Wallis test was employed. When significant differences were detected, pairwise comparisons were performed using the Dunn–Bonferroni post-hoc test. Descriptive values were presented as the median (min–max) due to the application of non-parametric tests. The intergroup comparison of categorical data was executed using the chi-square test, Fisher's exact test, and Fisher–Freeman–Halton test. Descriptive values for categorical data were expressed as the frequency (n) and percentage (%).

The significance level for all statistical analyses was set at α = 0.050. The SPSS version 25.0 (IBM Corp., Armonk, NY, USA) software package was utilized for data analysis.

Results

The indication for MRI examination was present for local staging of breast cancer in 13 patients, and in the remaining 203 patients, MRI was performed as part of a problem-solving approach for and screening of high-risk women. The lesions were evenly distributed between the right breast (51%) and the left breast (49%). The upper outer

quadrant was the most common location, accounting for 25% of the cases. The most common breast parenchymal pattern was type C (67%), and the most common type of background enhancement was minimal enhancement (53%) (Table 1).

In addition to detection by MRI, 18% of the lesions were also detected by MG and 39% by US. Follow-up was primarily performed using MRI in 76% (n = 164) of the cases, while the remaining 24% (n = 52) underwent follow-up with US and MG. Of the 164 lesions followed up with by MRI, 128 (78%) remained stable, and 36 (22%) disappeared. Among the patients monitored by US and MG (n = 52), 16 lesions disappeared, 31 remained stable, and 5 were determined to be benign (fibroadenoma, sclerosing adenosis, and intraductal papilloma) after surgical excision. The median follow-up interval was 18 months (range: 12–120 months).

The median lesion size measured by MRI was 8 mm (min: 3 mm; max: 25 mm). The following types of lesions were observed in MRI: 13.4% were classified as foci, 75.5% as masses, and 11.1% as non-mass enhancements (NMEs) (Table 1).

Among the most common types of masses, 98% were well circumscribed, with the predominant lesion shape being oval (57.1%) and the internal contrast pattern appearing as homogeneous in 68.1% of cases. In the masses, the kinetic curve type was persistent in 86% (type 1), the T2 signal was high in 50%, and the malignancy rate was 3.1%.

All foci were well circumscribed, with 79.3% showing homogeneous enhancement. Additionally, 90% showed no high signal on T2W images. The malignancy rate among the foci was 3.5%.

NMEs were primarily observed as focal contrast enhancements (62.5%), with the

majority exhibiting a homogeneous internal enhancement pattern (42%). No malignancy was detected among the NMEs, and the kinetic curve type was consistently persistent (92%), with no high signal observed on T2W images in 67% of cases.

When considering all lesions, the enhancement kinetic curve analysis showed that 87% of the lesions had a persistent enhancement, while 13% had a plateau enhancement (Table 1).

In 54% of the lesions, no distinct findings other than the parenchymal signal were detected on T2W images, while a high signal was observed in the same area as the lesion in 46%. Two patients exhibited a high T2 signal and received a malignant diagnosis; in both cases, the lesion type was identified as a mass.

An accompanying malignant mass was present in 13 patients, with invasive ductal carcinoma being the most common histopathological type (58%). All of the BI-RADS 3 lesions identified in these patients were classified in the benign category; among them, five were determined as benign through surgical excision, two were no longer apparent in follow-up MRI, and six remained stable over a median follow-up period of 34.5 months.

The follow-up results for the BI-RADS 3 lesions gave the following distribution of outcomes: 8% were excised, 0.5% decreased in size, 1.4% became enlarged, 17.1% disappeared, and 73% remained stable (Table 2).

Among the BI-RADS 3 lesions (n = 6) that were diagnosed as malignant, one lesion appeared as a focus, while five lesions showed mass enhancement in MRI. Among the malignant lesions categorized as masses, the shape was oval-lobulated in three patients and round in two patients. Regarding margin

features, two displayed irregular contours, while four exhibited smooth contours. In the internal contrast enhancement patterns, four presented a heterogeneous pattern, while two displayed a homogeneous pattern. In terms of kinetic curves, three lesions demonstrated a persistent pattern, and three followed a plateau pattern. Notably, four malignant lesions did not show a high T2 signal, whereas the other two exhibited high signal intensity.

Core needle biopsies were performed on 13 lesions; of those, 3 were diagnosed as malignant and 10 as benign. Two patients initially diagnosed with benign lesions were later found to have malignancy upon surgical excision (accuracy of biopsy: 84.6%, sensitivity: 100%, specificity: 60%, positive predictive value: 80%, negative predictive value: 100%).

Of the lesions that were surgically excised, 15 were determined to be benign, and 6 were diagnosed as malignant (Table 3). The malignancy upgrade rate was 2.8%.

Our further analysis of the clinical features of the BI-RADS 3 lesions diagnosed as malignant (Table 4) revealed the following findings: One patient presented with bloody nipple discharge during the one-year follow-up. Additionally, new suspicious microcalcifications were detected in two patients, and contour irregularity was observed in two patients on US control. In one patient, the follow-up of a lesion that initially displayed probably benign morphology on US was interrupted due to the novel coronavirus disease-2019 (COVID-19) pandemic. Two years later, the lesion exhibited both a significant increase in size and irregular contours (Figures 3, 4).

Table 1. Magnetic resonance imaging findings

		Follow-up results of the Breast Imaging Reporting and Data System (BI-RADS) 3 lesions		P
		Benign group n (%)	Malignant group n (%)	
Breast parenchymal pattern (n = 216)	**A**	11 (5.1)	0 (0)	
	B	43 (20)	2 (0.9)	
	C	140 (64.8)	4 (1.8)	0.838
	D	16 (7.4)	0 (0)	
Background enhancement (n = 216)	**Minimal**	113 (52.3)	2 (0.9)	
	Mild	24 (11.1)	0 (0)	
	Moderate	47 (21.8)	2 (0.9)	0.319
	Marked	26 (12.1)	(0.9)	
Lesion T2 signal (n = 216)	**Low and moderate**	112 (52)	98 (45.3)	
	High	4 (1.8)	2 (0.9)	0.688
Lesion type (n = 216)	**Focus**	28 (13)	(0.4)	
	Mass	158 (73.2)	5 (2.3)	1.000
	Non-mass	24 (11.1)	0 (0)	
Mass shape (n = 163)	**Oval**	92 (56.5)	1 (0.6)	0.020
	Lobulated	64 (39.3)	2 (1.2)	**Pairwise comparison test:** Lobulated–round P = 0.014 Oval–round P = 0.004
	Round	2 (1.2)	2 (1.2)	Oval–lobulated P = 0.570
Mass margin (n = 163)	**Smooth**	156 (95.8)	3 (1.8)	
	Irregular	2 (1.2)	2 (1.2)	0.004
Mass enhancement (n = 163)	**Homogeneous**	110 (67.5)	1 (0.6)	
	Heterogeneous	34 (20.9)	4 (2.4)	0.038 **Pairwise comparison test:**
	Rim	2 (1.2)	0 (0)	Homogeneous – heterogeneous P = 0.015
	Dark internal septation	12 (7.4)	0 (0)	
Non-mass enhancement (n = 24)	**Homogeneous**	10 (41.7)	0	
	Heterogeneous	9 (37.5)	0	-
	Clumped	5 (20.8)	0	
Kinetic curve type (n = 216)	**Persistent**	185 (88.1)	3 (50)	
	Plateau	25 (11.9)	3 (50)	0.030

Table 2. Follow-up results of the BI-RADS 3 lesions

Follow-up result	Stable n (%)	Decreased in size n (%)	Disappeared n (%)	Enlarged n (%)	Excised n (%)	Total n (%)
Benign group	158 (73.1)	1 (0.5)	37 (17.1)	2 (1)	12 (5.5)	210 (97.2)
Malignant group	-	-	-	1 (0.5)	5 (2.3)	6 (2.8)
Total	158 (73.1)	1 (0.5)	37 (17.1)	3 (1.4)	17 (7.8)	216 (100)
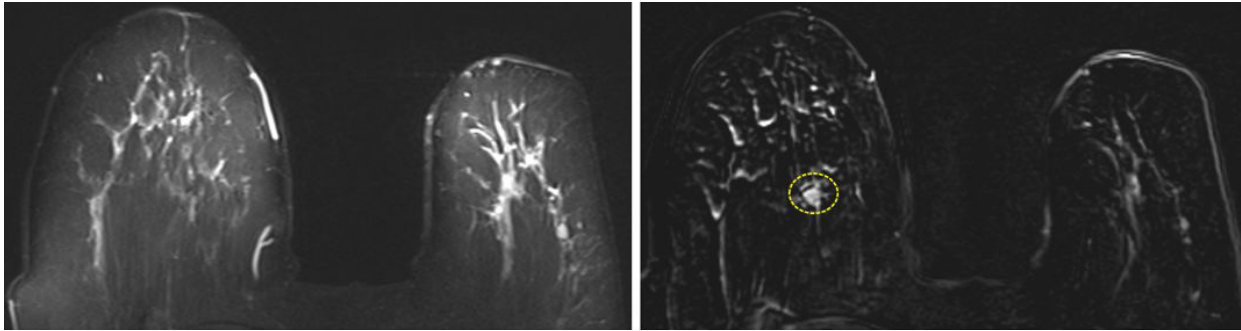
Figure 3. A 69-year-old female. No high signal was observed in the T2-weighted (T2W) image (right side). A heterogeneously enhanced 8-mm diameter mass was visible on the T1-weighted post-contrast subtraction magnetic resonance image (left side). It was categorized as a BI-RADS 3 lesion, and the patient was recommended for a follow-up magnetic resonance imaging after six months.
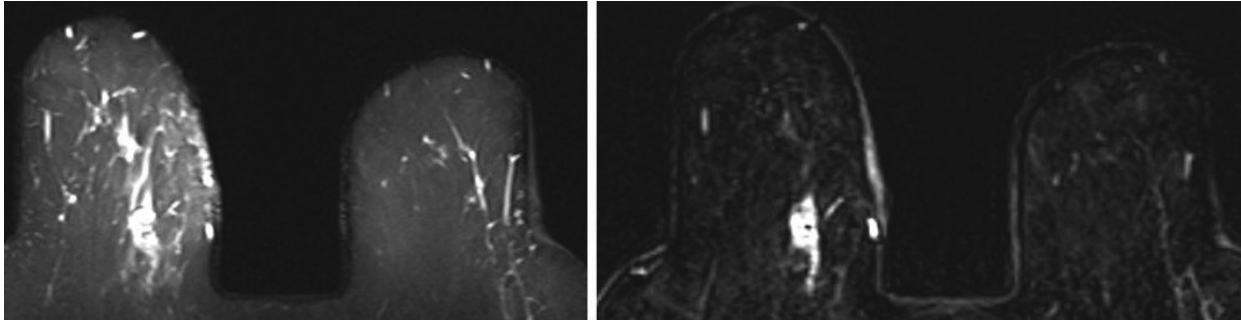
Figure 4. The patient depicted in **Figure 3** returned for a follow-up examination two years later, a delay attributed to the novel coronavirus-2019 pandemic. The lesion showed a substantial increase in size and displayed irregular contours on a T2-weighted image (right side) and a T1-weighted post-contrast subtraction magnetic resonance image (left side). The surgical excision histopathology result revealed an encapsulated papillary carcinoma.

Table 3. Histopathological results of the core needle biopsies and the surgical excisions

		n	%
Benign	Sclerosing adenosis	1	5.9
	Fibroadenoma	7	41.2
	Fibrocystic changes	1	5.9
	Atypical intraductal papilloma	2	11.7
Malignant	Invasive lobular carcinoma	1	5.9
	Ductal carcinoma in situ	3	17.6
	Invasive ductal carcinoma	1	5.9
	Encapsulated papillary carcinoma	1	5.9
	Total	17	100

When considering all lesion types, no statistically significant association was found between lesion localization, lesion diameter, T2 signal, and lesion type parameters for the possibility of upgrading to malignancy ($P > 0.050$). No significant relationship was observed between patient age and lesion upgrade to malignancy ($P = 0.084$).

A significant difference was evident between lesion shape and malignancy, as a progression to malignancy was more likely for round lesions than for lesions of other shapes ($P = 0.006$) (Table 1). An irregular margin, heterogeneous enhancement, and kinetic curve (type 2) features were significant for a lesion upgrade to malignancy ($P = 0.005$, $P = 0.015$, and $P = 0.030$, respectively) (Table 1).

An analysis of the relationship between the characteristics of masses and malignancy revealed significant associations between lesion shape, mass margins, internal enhancement patterns, and kinetic curve type ($P = 0.020$, $P = 0.004$, $P = 0.038$, and $P = 0.021$, respectively) (Table 1). No significant relationship was observed between the T2 signal and malignancy in masses ($P = 0.682$).

No malignancies were observed in NMEs; therefore, no statistical analysis of the variables could be conducted.

No significant differences were found between the lesion margin, internal enhancement pattern, and T2 signal in foci ($P > 0.050$).

Discussion

In this study, the malignancy rate of BI-RADS 3 lesions in MRI was 2.8%. The use of the BI-RADS 3 classification in MRI, despite the absence of specific morphological and kinetic features, results in a short interval follow-up in clinical practice.[3] Therefore BI-RADS 3 category in MRI assessment should be used precisely for increasing the sensitivity of the radiologist toward detecting early stage breast cancer and reducing the occurrence of unnecessary benign biopsies.[11] The prevalence of incidentally detected BI-RADS 3 lesions has increased due to the recommendation by the ACR for annual breast MRI control in patients with a lifetime risk of breast cancer above 20%.

The reported malignancy rate of BI-RADS 3 lesions seen in MRI varies widely, ranging from 0.5% to 10.1%.[3,4,6] One meta-analysis involving 2,183 lesions reported a malignancy upgrade rate of 50/2,183 and a notably higher malignancy rate in non-mass lesions.[4] Lourenco et al.[9] reported a malignancy rate of 2.4%, with the most common lesion form being a mass (49.1%) and the highest ma-

Table 4. Characteristics of the BI-RADS 3 lesions diagnosed as malignant at follow-up

Diameter of lesions (mm)	Reason for biopsy and surgical excision	Histopathological diagnosis
15	Bloody nipple discharge developing on follow-up	DCIS (low grade)
4	Suspicious microcalcifications added to the same area on follow-up	IDC
6	Mild contour irregularity developing on follow-up	IDC
8	Suspicious microcalcifications added to the same area on follow-up	ILC + LCIS
12	Enlargement of the nodule and subsequent structural distortion	IDC
8	Size increase and contour irregularity in the second year, which could not be followed up on due to the novel coronavirus-2019 pandemic	Encapsulated papillary carcinoma

DCIS, ductal carcinoma in situ; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; NAC, neoadjuvant chemotherapy; LCIS, lobular carcinoma in situ.

lignancy rate associated with NME being at 3.8%. Nguyen et al.[10], in their review, found a malignancy rate of 2% or lower for foci, 1.8% for mass lesions, and over 2% for NMEs. In our series, the malignancy upgrade rate was 2.8%, and the lesions diagnosed as malignant were predominantly in the form of mass enhancement. We identified a malignancy rate of 3.5% in foci and 3.1% in masses, whereas we did not detect any malignancy in lesions characterized by NMEs. The difference between our findings and the literature might reflect the fact that 75.5% of the 216 lesions in our study exhibited mass enhancement.

Regarding foci, Eby et al.[3] reported that lesions displaying persistent kinetic curves could potentially be benign. One study conducted by Ha et al.[13] on 136 foci indicated that kinetic evaluation was inconclusive. It further suggested that the absence of T2 hyperintensity and the presence of a newly developing or an enlarging focus would be a significant indicator of malignancy. In our study, 1 of the 29 foci was malignant, had no high T2 signal, showed heterogeneous enhancement, and had a persistent kinetic curve.

In certain studies, oval and round shapes have been grouped together for the assessment of malignancy. However, the review by Nguyen et al.[10] highlighted that a round shape holds significance in relation to malignancy. In our study, we identified a significant relationship between a round shape and malignancy ($P = 0.020$).

Our data were insufficient to draw conclusions concerning the relationship between lesion internal enhancement patterns and malignancy; therefore, considering this parameter as a factor in the decision-making process would be ill advised.[10] In the present study, we noted a significant distinction in the pairwise comparisons between the internal enhancement pattern of mass lesions and malignancy, particularly between homogeneous and heterogeneous types. Specifically, lesions with heterogeneous enhancement exhibited a higher malignancy rate ($P = 0.015$).

As malignancy was not identified among our patients with NMEs, we could not conduct any assessments in this regard. Grimm et al.[2] reported a greater prevalence of malignancy in lesions displaying NME and in those exhibiting inhomogeneous internal contrast enhancement. However, the data on NME kinetic properties are insufficient; therefore, we recommend looking at the distribution first.[10,11] The literature underscores that a linear or segmental distribution in lesions with NME is indicative of malignancy.[10,11,14] Among our NME lesions, only one displayed a linear distribution, while the rest were focal and regional. Importantly, all of these lesions were categorized as benign.

The Eastern Cooperative Oncology Group–ACR Imaging Network (ACRIN) 1141 trial advised classifying patients with and without lesions with hyperintense T2 signal as BI-RADS 3 if the lesions exhibited well-defined and homogeneous enhancement.[10,12] Even though the lesions with a high T2 signal were not included in the BI-RADS 3 category according to the latest ACR recommendation, lesions with a high T2 signal generally exhibit a lower malignancy rate.[10] Grimm et al.[15] demonstrated a malignancy rate below 2% for lesions exhibiting a high T2 signal. Conversely, Price et al.[16], who evaluated lesion characteristics within the BI-RADS 4 category, indicated that T2 hyperintensity is not a decisive characteristic. In our study, we did not identify any significant difference between the T2 signal and malignancy. The combined utilization of kinetic and morphological features provides support in the assessment for malignancy.[10]

Due to the high incidence of detecting additional cancers in newly diagnosed breast cancers, caution is needed when classifying lesions identified in MRI scans as BI-RADS 3 when the scans are performed for local staging.[17] The BI-RADS 3 category should not be applied in the absence of typical, likely benign findings or when lesion characterization cannot be executed.[11] Lee et al.[18] reported a malignancy rate of 3.5% for well-defined and rapidly enhancing lesions identified in MRI scans performed for local staging. Among the 13 MRI scans taken for local staging in the present study, all BI-RADS 3 lesions were categorized as benign and exhibited a persistent kinetic curve.

The malignancy rate in BI-RADS 3 lesions detected by MRI is generally low and falls within the accepted range for cancer rates observed in MG and US. Except for a malignancy in one patient who could not be followed up with due to the COVID-19 pandemic, the malignancies detected in the remaining five patients in our study were *in situ* and early stage cancers. The literature suggests that BI-RADS 3 lesions are commonly observed in the scans of high-risk patients and that their malignancy rates may be higher.[9,19] Our findings pointed to the importance of exercising caution during the follow-up of BI-RADS 3 lesions detected in high-risk patients and to the need for close monitoring for any new findings that may arise.

The utilization of second-look US in conjunction with MRI is a practical approach for the diagnosis and follow-up of lesions detected in MRI.[4,20,21] This method allows for further evaluation of the identified lesions, and subsequent follow-up can be performed using US.[10] Additionally, if necessary, core needle biopsy can be readily obtained under US guidance for histopathological diagnosis. In our study, lesions were identified during a second-look US in 84 patients. Among those patients, five individuals received a malignant diagnosis through biopsy and surgical excision, as suspicions escalated due to the margin characteristics of the lesions. Moreover, 37 patients exhibited probably benign findings, and they were subsequently monitored with ultrasonography (with a mean follow-up period of 33 months).

For the management of BI-RADS 3 lesions, which have variable malignancy rates, the literature recommends a total follow-up period of 24 months.[4,21] The initial follow-up

MRI examination is conducted within six months; if the lesion remains stable, the patient's follow-up continues, and a repeat MRI is performed after 12 months. If no concerning changes are noted, the lesion is downgraded to the BI-RADS 2 category.[10] However, any increase in size, morphological change, or appearance of additional suspicious findings during follow-up should be further evaluated with a biopsy.[21] One important point to note is that in the follow-up of lesions with a benign diagnosis from MRI-guided biopsies, approximately 8%–12% may have insufficient sampling, and 14%=18% may ultimately receive a malignant diagnosis; moreover, a false negative rate of 2.5% has been reported in MRI-guided biopsies.[21,22] These findings highlight the importance of careful monitoring and appropriate management of BI-RADS 3 lesions to ensure accurate diagnosis and timely intervention, if necessary.

Sadowski and Kelcz,[23] who studied 68 patients, identified four malignancies within a two-year follow-up period, leading to the recommendation of a comprehensive two-year follow-up interval. The ACRIN 6667 trial reported one malignancy (0.9%) among 106 patients during a two-year follow-up period, and this was attributed to the consistent experience level of the evaluating radiologist group.[24] In our study, we observed that the majority of patients with BI-RADS 3 lesions showed early stage *in situ* cancer during follow-up, with suspicious findings added at yearly intervals. We support the recommended follow-up intervals of six months, one year, and two years for BI-RADS 3 lesions detected in MRI, as suggested by the ACR.[1] Based on our results, we recommend a more careful evaluation in terms of malignancy, particularly for masses with a rounded shape, irregular contours, and heterogeneous enhancement. Considering the low malignancy rate of these lesions, shorter follow-up intervals in appropriate cases may help to reduce the number of unnecessary biopsies.

Our study had several limitations. One was that it was designed retrospectively, with a follow-up period of less than 24 months for some patients. Additionally, breast MRI scans were performed for different purposes, and the imaging quality varied due to the use of different MR devices (1.5 and 3 Tesla). Moreover, not all imaging modalities were available for all patients at their baseline and follow-up evaluations. Although some patients were not evaluated by breast MRI, they were assessed using other modalities, such as US or MG, providing consistency in the control evaluations. Diffusion-weighted MR images were not included in the study due to their unavailability for all patients. Since the BI-RADS 3 lesions were evaluated by a single radiologist, we were unable to evaluate interobserver consistency. Future prospective studies that examine a larger number of BI-RADS 3 lesions, include all imaging modalities, and use longer follow-up periods are needed to establish better clinical guidance and follow-up strategies.

In conclusion, the rising global utilization of breast MRI has created a demand for evidence-based standardized evaluation protocols for MRI BI-RADS 3 outcomes, similar to those established for MG and US. Assigning a classification of MRI BI-RADS 3 can pose challenges and might vary among radiologists and diagnostic centers. During follow-up, changes in size, morphology, and enhancement patterns are important potential indicators of malignancy development. Follow-up intervals should be determined on a case-by-case basis, taking these factors into account.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. American College of Radiology Breast Imaging Reporting and Data System (BI-RADS). 5th ed. Reston, VA: American College of Radiology; 2013. [CrossRef]

2. Grimm LJ, Anderson AL, Baker JA, et al. Frequency of malignancy and imaging characteristics of probably benign lesions seen at breast MRI. *AJR Am J Roentgenol.* 2015;205(2):442-447. [CrossRef]

3. Eby PR, DeMartini WB, Gutierrez RL, Saini MH, Peacock S, Lehman CD. Characteristics of probably benign breast MRI lesions. *AJR Am J Roentgenol.* 2009;193(3):861-867. [CrossRef]

4. Spick C, Bickel H, Polanec SH, Baltzer PA. Breast lesions classified as probably benign (BI-RADS 3) on magnetic resonance imaging: a systematic review and meta-analysis. *Eur Radiol.* 2018;28(5):1919-1928. [CrossRef]

5. Panigrahi B, Harvey SC, Mullen LA, et al. Characteristics and outcomes of BI-RADS 3 lesions on breast MRI. *Clin Breast Cancer.* 2019;19(1):e152-e159. [CrossRef]

6. Spick C, Szolar DH, Baltzer PA, et al. Rate of malignancy in MRI-detected probably benign (BI-RADS 3) lesions. *AJR Am J Roentgenol.* 2014;202(3):684-689. [CrossRef]

7. Hambly NM, Liberman L, Dershaw DD, Brennan S, Morris EA. Background parenchymal enhancement on baseline screening breast MRI: impact on biopsy rate and short-interval follow-up. *AJR Am J Roentgenol.* 2011;196(1):218-224. [CrossRef]

8. Aslan O, Oktay A, Yeniay L. Evaluation of the effects of neoadjuvant chemotherapy on probably benign breast lesions with MRI: report of two cases. *Ege Journal of Medicine.* 2022;61(2):298-301. [CrossRef]

9. Lourenco AP, Chung MT, Mainiero MB. Probably benign breast MRI lesions: frequency, lesion type, and rate of malignancy. *J Magn Reson Imaging.* 2014;39(4):789-794. [CrossRef]

10. Nguyen DL, Myers KS, Oluyemi E, et al. BI-RADS 3 assessment on MRI: a lesion-based review for breast radiologists [published correction appears in J Breast Imaging. 2023;5(3):e3-e12]. *J Breast Imaging* 2022;4(5):460-473. [CrossRef]

11. Lee KA, Talati N, Oudsema R, Steinberger S, Margolies LR. BI-RADS 3: Current and future use of probably benign. *Curr Radiol Rep.* 2018;6(2):5. [CrossRef]

12. Grimm LJ, Mango VL, Harvey JA, Plecha DM, Conant EF. Implementation of abbreviated breast MRI for screening: AJR Expert Panel Narrative Review. *AJR Am J Roentgenol.* 2022;218(2):202-212. [CrossRef]

13. Ha R, Sung J, Lee C, Comstock C, Wynn R, Morris E. Characteristics and outcome of enhancing foci followed on breast MRI with management implications. *Clin Radiol.* 2014;69(7):715-720. [CrossRef]

14. Kim JA, Son EJ, Youk JH, et al. MRI findings of pure ductal carcinoma in situ: kinetic characteristics compared according to lesion type and histopathologic factors. *AJR Am J Roentgenol.* 2011;196(6):1450-1456. [CrossRef]

15. Grimm LJ, Enslow M, Ghate SV. Solitary, Well-circumscribed, T2 hyperintense masses on MRI have very low malignancy rates. *Journal of Breast Imaging.* 2019;1(1):37-42. [CrossRef]

16. Price ER, Sickles EA, Yitta S, et al. Use of the probably benign (BI-RADS category 3) assessment for masses on breast MRI: Is it transferable to general clinical practice? *Breast J.* 2018;24(2):109-114. [CrossRef]

17. Liberman L, Morris EA, Kim CM, et al. MR imaging findings in the contralateral breast of women with recently diagnosed breast cancer. *AJR Am J Roentgenol.* 2003;180(2):333-341. [CrossRef]

18. Lee JY, Jang M, Kim SM, Yun B, Jang JY, Ahn HS. Preoperative magnetic resonance imaging characteristics of oval circumscribed fast enhancing lesions in patients with newly diagnosed breast cancer. *Medicine (Baltimore).* 2018;97(19):e0704. [CrossRef]

19. Liberman L, Morris EA, Benton CL, Abramson AF, Dershaw DD. Probably benign lesions at breast magnetic resonance imaging: preliminary experience in high-risk women. *Cancer.* 2003;98(2):377-388. [CrossRef]

20. Spick C, Baltzer PA. Diagnostic utility of second-look US for breast lesions identified at MR imaging: systematic review and meta-analysis. *Radiology*. 2014;273(2):401-409. [CrossRef]

21. Lee KA. Breast Imaging Reporting and Data System category 3 for magnetic resonance imaging. *Top Magn Reson Imaging*. 2014;23(6):337-344. [CrossRef]

22. Li J, Dershaw DD, Lee CH, Kaplan J, Morris EA. MRI follow-up after concordant, histologically benign diagnosis of breast lesions sampled by MRI-guided biopsy. *AJR Am J Roentgenol*. 2009;193(3):850-855. [CrossRef]

23. Sadowski EA, Kelcz F. Frequency of malignancy in lesions classified as probably benign after dynamic contrast-enhanced breast MRI examination. *J Magn Reson Imaging*. 2005;21(5):556-564. [CrossRef]

24. Weinstein SP, Hanna LG, Gatsonis C, Schnall MD, Rosen MA, Lehman CD. Frequency of malignancy seen in probably benign lesions at contrast-enhanced breast MR imaging: findings from ACRIN 6667. *Radiology*. 2010;255(3):731-737. [CrossRef]

Long-term results of liver thermal ablation in patients with hepatocellular carcinoma and colorectal cancer liver metastasis regarding spatial features and tumor-specific variables

Okan Akhan[1]

Ahmet Gürkan Erdemir[1]

Sevilay Karahan[2]

Emre Ünal[1]

Türkmen Turan Çiftçi[1]

Devrim Akıncı[1]

Şuayib Yalçın[3]

[1]Hacettepe University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

[2]Hacettepe University Faculty of Medicine, Department of Biostatistics, Ankara, Türkiye

[3]Hacettepe University Faculty of Medicine, Department of Medical Oncology, Ankara, Türkiye

PURPOSE

Colorectal cancer liver metastasis (CRLM) and hepatocellular carcinoma (HCC) are widely treated using microwave and radiofrequency ablation. Local tumor progression (LTP) may develop depending on the shortest vascular distance and large lesion diameter. This study aims to explore the effect of these spatial features and to investigate the correlation between tumor-specific variables and LTP.

METHODS

This is a retrospective study covering the period between January 2007 and January 2019. One hundred twenty-five patients (CRLM: HCC: 64:61) with 262 lesions (CRLM: HCC: 142:120) were enrolled. The correlation between LTP and the variables was analyzed using the chi-square test, Fischer's exact test, or the Fisher-Freeman-Halton test where applicable. The local progression-free survival (Loc-PFS) was analyzed using the Kaplan–Meier method. Univariable and multivariable Cox regression analyses were performed to identify prognostic factors.

RESULTS

Significant correlations were observed for LTP in both CRLM and HCC at a lesion diameter of 30–50 mm ($P = 0.019$ and $P < 0.001$, respectively) and SVD of ≤3 mm ($P < 0.001$ for both). No correlation was found between the ablation type and LTP (CRLM: $P = 0.141$; HCC: $P = 0.771$). There was no relationship between residue and the ablation type, but a strong correlation with tumor size was observed ($P = 0.127$ and $P < 0.001$, respectively). In CRLM, LTP was associated with mutant K-ras and concomitant lung metastasis ($P < 0.001$ and $P = 0.003$, respectively). In HCC, a similar correlation was found for Child–Pugh B, serum alpha-fetoprotein (AFP) level of >10 ng/mL, predisposing factors, and moderate histopathological differentiation ($P < 0.001$, $P = 0.008$, $P = 0.027$, and $P<0.001$, respectively). In CRLM, SVD of ≤3 mm proved to be the variable with the greatest negative effect on Loc-PFS ($P = 0.007$), followed by concomitant lung metastasis ($P = 0.027$). In HCC, a serum AFP level of >10 ng/mL proved to be the variable with the greatest negative effect on Loc-PFS ($P = 0.045$).

CONCLUSION

In addition to the lesions' spatial features, tumor-specific variables may also have an impact on LTP.

KEYWORDS

Ablation techniques (D055011), colorectal neoplasm (D015179), hepatocellular carcinoma (D006528), local tumor progression (D009364), survival analysis (D016019)

Corresponding author: Okan Akhan

E-mail: drokanakhan@gmail.com, akhano@tr.net

Colorectal cancer liver metastasis (CRLM) and hepatocellular carcinoma (HCC) are widely treated using local ablation, which provides an increased survival outcome.[1-3] Both radiofrequency ablation (RFA) and microwave ablation (MWA), the most commonly used ablation techniques, cause necrosis through elevated temperatures, but they have different physical parameters that offer advantages in certain situations.[4,5]

After local ablation procedures, local tumor progression (LTP) may be confronted out of favor.[6,7] This situation is more frequent in large tumors that exceed the perimeter of the

ablation zone and in tumors with a blood vessel in close proximity (heat-sink effect).[8,9] Of these two variables, tumors with a blood vessel in close proximity are considered the highest risk factor for the development of LTP.[5,10,11] Given the extensive literature addressing the development of LTP, it is possible that several non-spatial variables also influence this development in malignancies with different pathogenesis, such as mutant K-ras oncogene in colorectal carcinoma (CRC) and alpha-fetoprotein (AFP) levels before ablation in HCC.

This study has three main objectives. First, to investigate the correlation between common variables (tumor diameter, shortest vascular distance, and the ablation type) and LTP with local progression-free survival (Loc-PFS), which corresponds to the period without LTP. Second, to investigate the association with K-ras mutation, primary tumor location (left- or right-sided), and concomitant lung metastases in CRLM as tumor-specific variables. Third, to investigate the association between the Child–Pugh score, histopathological differentiation grade, serum AFP level, and predisposing factors for chronic liver disease in HCC as tumor-specific variables.

Methods

Study design

This study is a retrospective analysis of liver lesions that received RFA or MWA between January 2007 and January 2019 due to CRLM or HCC. The Hacettepe University Faculty of Medicine Ethics Committee of the institute approved this study (GO-18/429).

The decision for each thermal ablation was made by the multidisciplinary Institutional tumor board, and informed consent was obtained from all enrolled patients.

Imaging-guided ablation therapies were defined according to publications developed

Main points

- Morphometric features of the lesion, such as larger diameters or shorter vascular proximity, were an effective factor in local tumor progression.

- Colorectal cancer liver metastasis, concomitant lung metastasis, and a host-specific variable had the greatest impact on local progression-free survival after short vascular proximity.

- In hepatocellular carcinoma, a serum alpha-fetoprotein level of >10 ng/mL proved to be the variable with the greatest negative effect on local progression-free survival.

by the "International Working Group on Image-Guided Tumor Ablation" and "Results of the SIO and DATECAN Initiative".[12,13]

Inclusion criteria

The inclusion criteria for both CRLM and HCC lesions were as follows:

1. Maximum of five liver lesions for each patient with CRC and a maximum of three lesions for each patient with HCC,

2. Maximum diameter of 5 cm for each lesion,

3. Curative intent (the ablation of all liver lesions in the same session),

4. Presence of magnetic resonance imaging (MRI) images within 2 months before the ablation,

5. Presence of follow-up MRI or computed tomography (CT) imaging at 1, 3, 6, and 12 months and semiannually after the first year.

Furthermore, the presence of lung metastasis was not an exclusion criterion.

Seventy-one patients with CRC and 67 patients with HCC who underwent ablation with "curative intent" were identified. However, due to insufficient follow-up, seven patients with CRC (9.85%) and six patients with HCC (8.95%) were excluded. Finally, the remaining 64 patients with CRC (142 lesions) and 61 patients with HCC (120 lesions) were enrolled in this study.

For further information please see the flowchart (Figure 1).

Ablation procedure and follow-up

All procedures were done with ultrasound guidance. StarBurst® (AngioDynamics®) electrodes were used for RFA, while Acculis®/Solero® (AngioDynamics®) antennas were used for MWA. All procedures were performed according to the manufacturer's instructions, with an ablation margin of at least 5–10 mm.[14]

After the ablation, contrast-enhanced CT and MRI were performed within the first month. Patients with no residual disease were accepted as "complete ablation". Tumor development during follow-up in patients with complete ablation was classified as LTP. Loc-PFS was calculated for each lesion, starting with ablation until the development of LTP or patient death, and it was censored at the last follow-up date. The censor date for Loc-PFS estimation was February 2, 2020.

Data collection

All data were reviewed and collected with the consensus of two radiologists (A.G.E. and O.A.) at two different time points to ensure external and internal validity in both patient selection and data collection.

The segmental distribution, LTP development, thermal ablation type (MWA or RFA), and shortest vascular distance of each lesion were recorded. The measurement of the shortest vascular distance was performed on volumetric dynamic T1W slices from the patient's last MRI before ablation. The shortest perpendicular distance to the vessel with a width of ≥3 mm was estimated through multiplanar reformation images (Figure 2). The longest axial and craniocaudal diameter of each lesion was also recorded.

The complications, ablation type and technique (percutaneous or intraoperative), and the segment of the relevant lesion were recorded.

The presence of K-ras mutation, the site of primary disease (right or left colon), and concomitant lung metastasis were considered CRC-specific variables. In the HCC group, predisposing factors (non-alcoholic steatohepatitis and hepatitis B or C virus), degree of histopathologic differentiation, AFP level, and Child–Pugh score within one month before ablation were recorded.

Histopathologic diagnoses were available in both the CRC and HCC groups that participated in this study. However, in patients with CRC with multiple liver metastases, only one of the lesions was biopsied. In addition, K-ras mutations were analyzed using DNA derived from formalin-fixed paraffin-embedded tumors obtained from primary sites in the colon. The presence of microsatellite instability was also analyzed.

Statistical analysis

The data were processed using the IBM-SPSS® Statistics 24.0, StataCorp LCC-STATA® 14 software, and R® version 4.0.3. Categorical variables were reported as frequencies and percentages, and continuous variables were reported as means and standard deviations.

Categorical variables were evaluated using the chi-square test or the Fisher-Freeman-Halton test where applicable. For all tests, a two-tailed P value of less than 0.05 was considered statistically significant.

The Kaplan–Meier method was used for estimates of Loc-PFS, and the log-rank test was used to compare survival groups. Cox

Figure 1. The flowchart of patient selection. CRC, colorectal carcinoma; HCC, hepatocellular carcinoma; MWA, microwave ablation; RFA, radiofrequency ablation.

Results

Background data and complications

The detailed baseline characteristics of 262 lesions and 125 patients are shown with all aspects in Table 1.

The complications of the included patients were biliary obstruction, abscess, and costochondritis (Figures 3-5). Seventeen lesions (out of 262 lesions, 6.48%) were complicated: eight of them were percutaneous (out of 198 lesions, 4.04%) and nine of them were intraoperative (out of 64 lesions, 14.06%). A significant correlation was found between intraoperative ablation and the occurrence of complications: when all 17 complications were included and when only the abscess [three percutaneous (3/198 = 1.51%) and six intra-operative (6/64 = 9.37%) lesions] were included ($P < 0.005$). Six of the nine abscesses had a history of hepaticojejunostomy (two lesions) and endoscopic sphincterotomy (four lesions) due to gallstones. All lesions (n = 7, 2.67%) that developed biliary dilatation were in the central segments (segments 1, 4b, and 5) ($P < 0.001$). A transient costochondritis complication was observed in only one patient with a subcapsular localized lesion in segment eight.

regression models were used to assess the effects of confounding factors on overall survival. Variables with a P value of <0.20 in the univariable analyses were analyzed in multivariable Cox regression models to explore prognostic factors of overall survival. The results are reported with hazard ratios and 95% confidence intervals.[15]

Table 1. General background data						
	Colorectal cancer		Hepatocellular carcinoma		Total	
	RFA	MWA	RFA	MWA	RFA	MWA
Patients* (n)	31 (48.73%)	33 (51.57%)	35 (57.37%)	26 (42.63%)	66 (52.80%)	59 (47.20%)
Gender (M:F)	20:11	20:13	22:13	17:9	42:24	37:21
Age†	57.74 ± 11.88	61.25 ± 5.58	61.88 ± 9.62	61.20 ± 11.03	59.15 ± 10.93	61.24 ± 9.85
Lesions* (n)	57 (40.14%)	85 (59.86%)	71 (52.98%)	49 (47.02%)	128 (48.85%)	134 (51.15%)
<30 mm (n)	42 (36.52%)	73 (63.48%)	60 (61.22%)	38 (38.78%)	102 (47.88%)	111 (52.12%)
30–50 mm (n)	15 (55.55%)	12 (44.45%)	11 (50.00%)	11 (50.00%)	26 (53.06%)	23 (46.94%)
R-L diameter†	17.87 ± 7.76	20.42 ± 7.80	16.78 ± 7.42	18.67 ± 9.37	17.11 ± 7.71	20.11 ± 9.01
A-P diameter†	16.87 ± 7.31	19.52 ± 7.60	16.71 ± 7.22	17.91 ± 8.48	16.75 ± 7.24	18.84 ± 8.20
C-C diameter†	16.95 ± 7.72	19.85 ± 7.96	16.94 ± 7.43	18.69 ± 9.29	16.94 ± 7.52	19.18 ± 8.44
Segmental distribution* (n)						
Segment-1	1 (33.33%)	2 (66.66%)	2 (100.0%)	0 (0.00%)	3 (60.00%)	2 (40.00%)
Segment-2	4 (30.76%)	9 (69.24%)	4 (57.14%)	3 (42.86%)	8 (40.00%)	12 (60.00%)
Segment-3	3 (42.85%)	4 (57.15%)	6 (75.00%)	2 (25.00%)	9 (60.00%)	6 (40.00%)
Segment-4a	9 (47.36%)	10 (52.64%)	8 (53.33%)	7 (46.67%)	17 (50.00%)	17 (50%)
Segment-4b	3 (50.00%)	3 (50.00%)	3 (75.00%)	1 (25.00%)	6 (60.00%)	4 (40%)
Segment-5	9 (40.90%)	13 (59.10%)	17 (73.91%)	6 (26.09%)	26 (57.77%)	19 (42.23%)
Segment-6	9 (40.90%)	13 (59.10%)	11 (55.00%)	9 (45.00%)	20 (47.61%)	22 (52.39%)
Segment-7	8 (40.00%)	12 (60.00%)	6 (40.00%)	9 (60.00%)	14 (40.00%)	21 (60.00%)
Segment-8	11 (36.66%)	19 (63.34%)	14 (53.84%)	12 (46.16%)	25 (44.64%)	31 (55.36%)
The shortest vascular distance* (n)						
≤3 mm (n)	14 (43.75%)	18 (56.25%)	11 (55.00%)	9 (45.00%)	25 (48.07%)	27 (51.93%)
>3 mm (n)	43 (39.09%)	67 (60.91%)	60 (60.00%)	40 (40.00%)	103 (49.04%)	107 (50.96%)

†Mean values are given as millimeters with their ± standard deviations. *The percentages in parentheses show the individual distributions of frequencies within the CRC, HCC, and Total groups, depending on which ablation technique was chosen. A-P, anterior-posterior diameter; C-C, craniocaudal diameter; CRC, colorectal carcinoma; HCC, hepatocellular carcinoma; R-L, right-left diameter; RFA, radiofrequency ablation; MWA, microwave ablation.
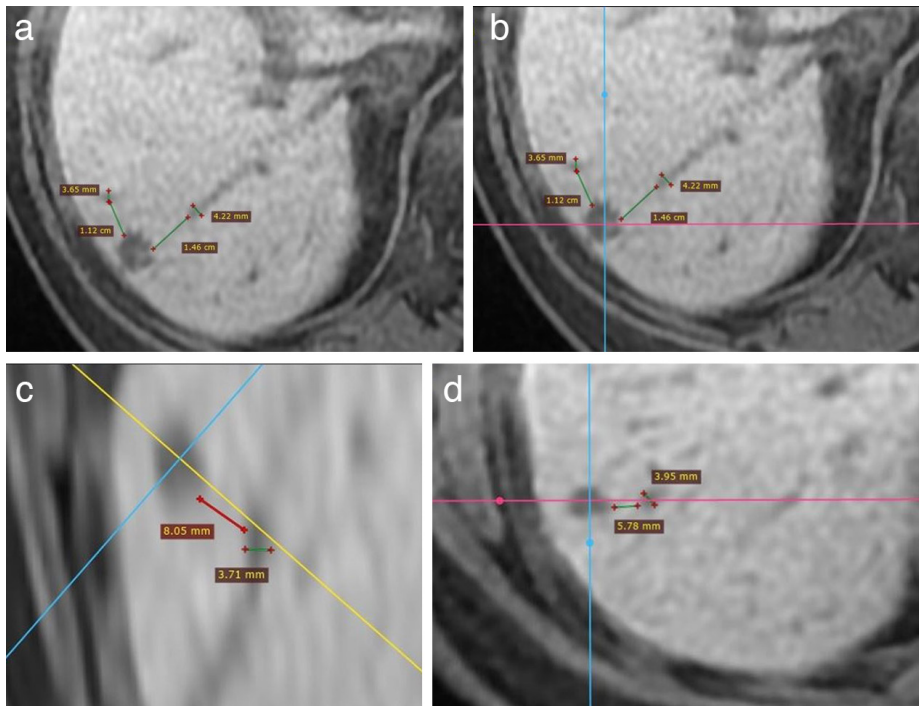
Figure 2. The measurement of the shortest vascular distance. Dynamic T1W volume sections in a patient with CRC metastasis in segment 6 are shown **(a)**. Two vessels with the smallest distance to the lesion, 1.12 mm and 1.46 mm and a width of approximately 4 mm, (3.65 mm and 4.22 mm, respectively) are seen in the axial sections. To determine the exact distance, the dimensional indicators were centered on the lesion **(b)**. Rotating through 360 degrees in the coronal and sagittal planes **(c)**, the closest vessel distance was sought. On this plane represented with the yellow line **(c)**, the exact distance was determined to be 5.78 mm in the axial-oblique section **(d)**. CRC, colorectal carcinoma.
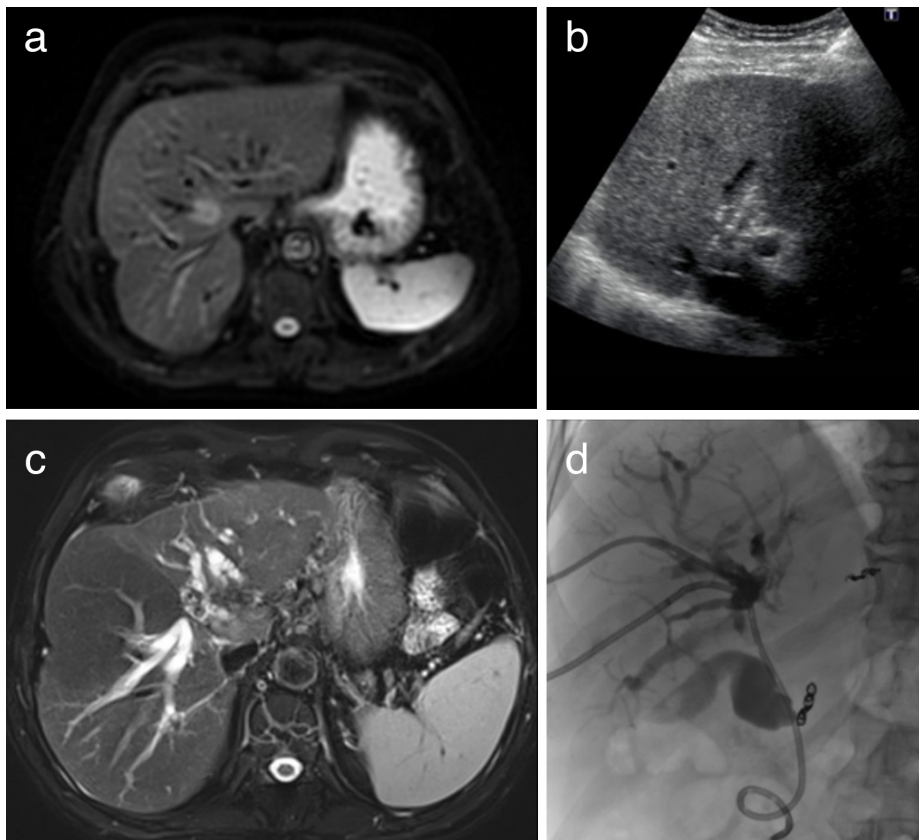
Figure 3. Biliary obstruction after RFA of CRC metastasis. A segment 4b metastasis is seen on the fat-suppressed T2-weighted slice **(a)**. Ultrasound-guided percutaneous radiofrequency ablation is performed **(b)**. Approximately nine months after the procedure, the patient developed biliary dilatation **(c)** due to the central ablation scar, and percutaneous biliary drainage **(d)** was performed. RFA, radiofrequency ablation; CRC, colorectal carcinoma.

Residue occurred in three CRLM (MWA: RFA: 0:3) and four (MWA: RFA: 1:3) HCC lesions, and six of seven residual lesions (out of 262 lesions) were observed after RFA (P = 0.127). Moreover, all six of them had a diameter of 30–50 mm ($P < 0.001$). For all lesions with residual occurrences, that were reablated with complete ablation, were included in the cohort from the time of complete ablation.

More detailed information on complications and residue is shown in Table 2.

Correlations of common variables with local tumor progression development and local progression-free survival

Regarding the ablation type (MWA or RFA), no statistically significant difference was found for LTP development and Loc-PFS in the CRC group ($P = 0.141$ and $P = 0.161$, respectively). In the HCC group, no correlation was found between the development of LTP and Loc-PFS considering the ablation type ($P = 0.771$ and $P = 0.699$, respectively).

The development rate of LTP in CRLM was statistically significant in those with a lesion diameter of 30–50 mm ($P = 0.019$). Loc-PFS also decreased in this group but failed to reach a statistically significant result ($P = 0.085$). In HCC lesions with a lesion diameter of 30–50 mm, a statistically significant correlation was observed between both LTP development and Loc-PFS ($P < 0.001$ for both).

The shortest vascular distance of ≤3 mm in both HCC and CRLM was statistically associated with both LTP development ($P < 0.001$ for each group) and decreased Loc-PFS ($P < 0.001$ and $P = 0.014$, respectively).

More detailed information on the common variables of both groups can be found in Tables 3 and 4.

Multivariable analysis and correlations of colorectal carcinoma-specific variables with local tumor progression and local progression-free survival

In the Cox regression analysis for the CRC-specific variable model (Supplementary Table 1), the P value was 0.0006.

Mutated K-ras oncogene was found to be statistically correlated with both LTP development and decreased Loc-PFS ($P < 0.001$ and $P = 0.021$, respectively). Similar results were observed with the existence of concomitant lung metastasis for both LTP development and decreased Loc-PFS ($P = 0.003$ and $P = 0.044$, respectively). Although LTP

Table 2. Residue, complications, and puncture type

	Colorectal cancer		Hepatocellular carcinoma		Total	
	RFA	MWA	RFA	MWA	RFA	MWA
Percutaneous thermoablation†						
Patients (n)	25 (56.81%)	19 (43.19%)	28 (57.14%)	21 (42.86%)	53 (56.98%)	40 (43.02%)
Lesions (n)	49 (49.49%)	50 (50.51%)	58 (58.58%)	41 (41.42%)	107 (54.04%)	91 (45.96%)
Complications						
Biliary dilatation						
Patients (n)	2	-	2	-	4	-
Lesions (n)	2	-	2	-	4	-
Localization	Segment-4b **(1)**	-	Segment-5 **(2)**	-	Segment-4b **(1)**	-
	Segment-5 **(1)**				Segment-5 **(4)**	
Abscess						
Patients (n)	-	-	3	-	3	-
Lesions (n)	-	-	3	-	3	-
Localization	-	-	Segment-6 **(2)**	-	Segment-6 **(2)**	-
			Segment-8 **(1)**		Segment-8 **(1)**	
Costochondritis						
Patients (n)	-	-	1	-	1	-
Lesions (n)	-	-	1	-	1	-
Localization	-	-	Segment-8 **(1)**	-	Segment-8 **(1)**	-
Intra-operative thermoablation†						
Patients (n)	6 (30.00%)	14 (70.00%)	7 (58.33%)	5 (41.67%)	13 (40.62%)	19 (59.38%)
Lesions (n)	8 (18.60%)	35 (81.40%)	13 (61.90%)	8 (38.10%)	21 (32.81%)	43 (67.19%)
Complications						
Biliary dilatation						
Patients (n)	-	2	1	-	1	2
Lesions (n)	-	2	1	-	1	2
Localization	-	Segment-1 **(1)**	Segment-4b **(1)**	-	Segment-4b **(1)**	Segment-1 **(1)**
		Segment-5 **(1)**				Segment-5 **(1)**
Abscess						
Patients (n)	1	3	1	1	2	4
Lesions (n)	1	3	1	1	2	4
Localization	Segment-5 **(1)**	Segment-5 **(1)**	Segment-5 **(1)**	Segment-4a **(1)**	Segment-5 **(2)**	Segment-4a **(1)**
		Segment-7 **(2)**				Segment-5 **(1)**
						Segment-7 **(2)**
Residue*						
Patients (n)	3	-	3	1	6	1
Lesions (n)	3	-	3	1	6	1
Lesion diameter (n) (m)	3	-	3	0	6	0
Shortest vascular distance (n) (≤3 mm)	0	-	0	1	0	1

†The percentages in parentheses after the frequency values show individual distributions within groups (CRC, HCC, and total), depending on which ablation technique was chosen. Bold parentheses after the segments show the segmental distribution. *For all lesions with residual occurrences that were reablated with complete ablation and were included in the cohort from the time of complete ablation. CRC, colorectal carcinoma; HCC, hepatocellular carcinoma; RFA, radiofrequency ablation; MWA, microwave ablation.

development and decreased Loc-PFS were more associated with right-sided CRLM, no statistically significant results were observed ($P = 0.064$ and $P = 0.358$, respectively).

Although microsatellite instability was also analyzed in all patients, it was not detected in any of them.

In the multivariable analysis for CRLMs, the shortest vascular distance of ≤3 mm was found to be the variable with the largest negative effect on Loc-PFS ($P = 0.007$), followed by concomitant lung metastasis ($P = 0.027$).

More detailed information on CRC-specific variables and multivariable analysis can be found in Table 3 and Supplementary Table 1.

Multivariable analysis and correlations of hepatocellular carcinoma-specific variables with local tumor progression and local progression-free survival

In the Cox regression analysis for the HCC-specific variable model (Supplementary Table 1), the P value was 0.0002.

Child–Pugh B, a serum AFP level of >10 ng/mL, and moderate histopathological differentiation showed a highly significant statistical correlation with both LTP development and decreased Loc-PFS ($P < 0.001$, $P = 0.008$, and $P < 0.001$, respectively).

Poor histopathologic differentiation was not observed in the entire HCC cohort.

The LTP development rate in lesions with HBV was statistically significant ($P = 0.027$). However, although Loc-PFS decreased in this group, no statistically significant results were obtained ($P = 0.210$).
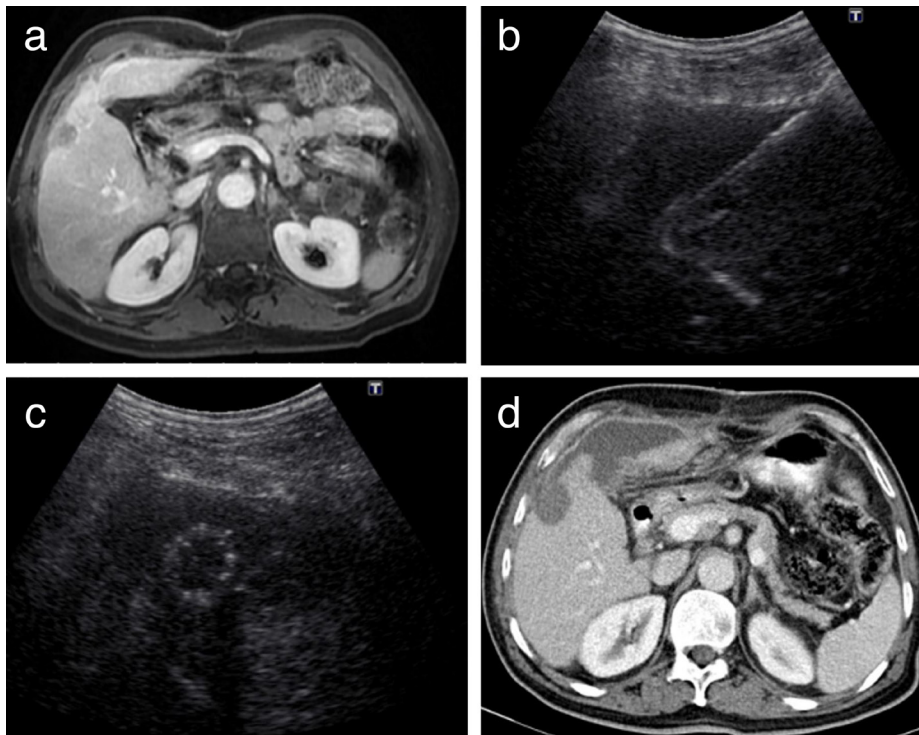
Figure 4. Abscess formation after MWA of CRC metastasis. A segment-5 metastasis is visible on the portal venous phase enhanced MRI **(a)**. Ultrasound-guided intraoperative MWA and the ablation zone with echogenic borders are seen **(b, c)**. On the fourth day after surgery, an abscess associated with the ablation zone and subcapsular suppuration were observed on contrast-enhanced abdominal CT examination **(d)**, which was performed after the addition of fever to persistent right upper quadrant pain. CRC, colorectal carcinoma; MWA, microwave ablation; CT, computed tomography; MRI, magnetic resonance imaging.
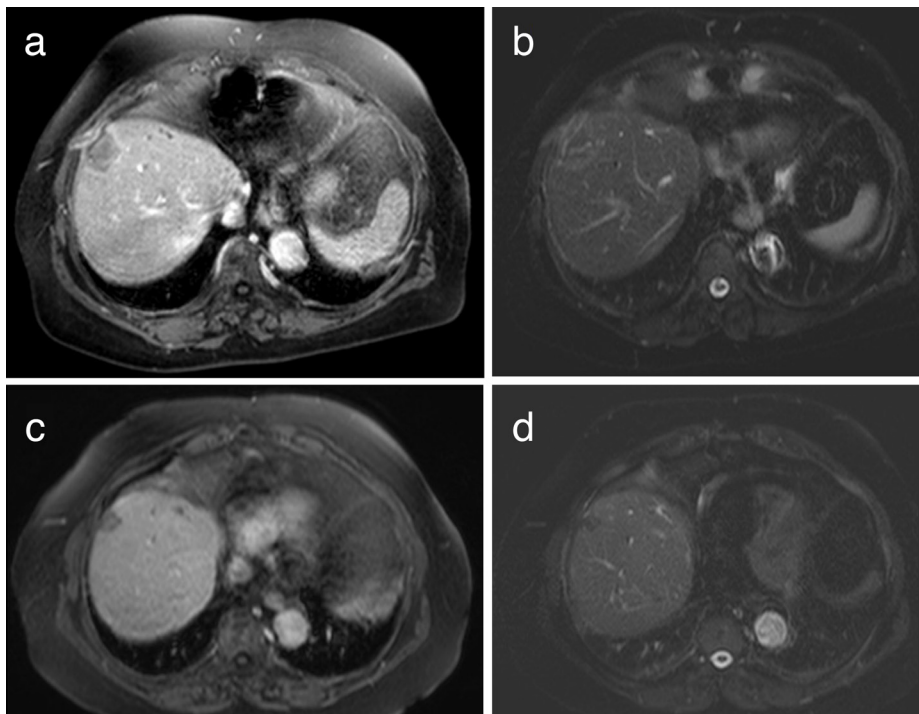
Figure 5. Costochondritis after RFA of HCC. A patient with segment 8 HCC who underwent RFA one month ago has right upper quadrant pain that does not resolve. Follow-up MRI in the first month shows a costochondral inflammatory signal increase adjacent to the ablation zone in the postcontrast T1 **(a)** and fat-suppressed T2 slices **(b)**. At the sixth month follow-up, the postcontrast fat-suppressed T1 **(c)** and fat-suppressed T2 **(d)** slices show a regression of costochondral inflammation and a shrunken ablation cavity. RFA, radiofrequency ablation; HCC, hepatocellular carcinoma; MRI, magnetic resonance imaging.

In the multivariable analysis for HCC lesions, a serum AFP level of >10 ng/mL was found to be the variable with the largest negative effect on Loc-PFS ($P = 0.045$).

More detailed information on HCC-specific variables and multivariable analysis can be found in Table 4 and Supplementary Table 1.

Discussion

This study's results showed that in both CRLM and HCC lesions, albeit with lower LTP rates in the lesions treated with MWA, there was no significant difference between RFA and MWA. In contrast, two other common variables in both lesion groups were statistically associated with LTP: the shortest vascular distance of ≤3 mm and a lesion diameter of 30–50 mm. Important results were obtained from the observations for tumor-specific variables. A significant correlation of CRC-specific variables was observed with mutant K-ras and concomitant lung metastasis, while the same was observed for HCC-specific variables with Child–Pugh B, a serum AFP level of >10 ng/mL, HBV, and moderate histopathological differentiation.

Extensive meta-analyses have shown that the most important difference in the clinical outcome between MWA and RFA is the size of the larger liver lesion treated, with RFA having some possible disadvantages over LTP.[9,16,17] In this study, although residues were seen more frequently in tumors treated with RFA, all of these lesions were 30–50 mm in diameter. Although no statistically significant results were obtained in this cohort, LTP was more frequent, and Loc-PFS was shorter in patients treated with RFA. In accordance with this study, numerous articles have been published in the literature showing the association of tumor size and shortest vascular distance with LTP.[7,11,18-20]

Recent retrospective studies have shown a strong correlation between the K-ras mutation, which is one of the CRC-specific variants, and LTP.[21,22] In the study by Jiang et al.[23], which is one of the most recent studies conducted in this context, similar results were obtained, but they only included lesions with RFA. The second tumor-specific variable studied in CRLM lesions was the primary origin of the tumor. There are few studies in the literature that address primary origins. Zhou et al.[24], who studied patients with MWA, and Gu et al.[25], who studied patients with RFA, conduct-

Table 3. LTP development and Loc-PFS of CRLMs

	LTP development			Loc-PFS				
	Developed	Not developed	P value*	1-year survival	3-year survival	5-year survival	Median survival (months)	P value*
Ablation type								
MWA	15 (17.64%)	70 (82.36%)	0.141	80.00% (n = 64)	43.52% (n = 37)	29.41% (n = 25)	57.26	0.161
RFA	16 (28.07%)	41 (71.93%)		73.68% (n = 42)	68.42% (n = 39)	68.42% (n = 39)	98.31	
Lesion diameter								
<30 mm	19 (16.52%)	96 (83.48%)	0.019	74.78% (n = 86)	60.00% (n = 69)	49.56% (n = 57)	78.41	0.085
3–50 mm	12 (44.44%)	15 (55.55%)		74.07% (n = 20)	25.92% (n = 7)	25.92% (n = 7)	45.74	
The shortest vascular distance								
≤3 mm	23 (71.87%)	9 (28.13%)	<0.001	62.50% (n = 20)	37.50% (n = 12)	15.62% (n = 5)	8.06	<0.001
>3 mm	8 (7.27%)	102 (92.73%)		78.18% (n = 86)	58.18% (n = 64)	53.63% (n = 59)	78.11	
K-ras oncogene								
Wild	5 (6.75%)	69 (93.25%)	<0.001	83.78% (n = 62)	77.02% (n = 57)	77.02% (n = 57)	47.02	0.044
Mutated	26 (38.23%)	42 (61.77%)		61.76% (n = 42)	27.94% (n = 19)	10.29% (n = 7)	105.71	
Right/left sided								
Right colon	10 (34.48%)	19 (65.52%)	0.064	75.86% (n = 22)	48.27% (n = 14)	31.03% (n = 9)	40.77	0.358
Left colon	21 (18.58%)	92 (81.42%)		74.33 (n = 84)	54.86% (n = 62)	48.67% (n = 55)	87.13	
Concomitant lung metastasis								
Yes	14 (40.00%)	21 (60.00%)	0.003	42.85% (n = 15)	25.71% (n = 9)	25.71% (n = 9)	22.32	0.021
No	17 (15.88%)	90 (84.12%)		77.77% (n = 91)	57.26% (n = 67)	47.01% (n = 55)	87.11	

*These P values indicate the results of univariate Cox regression analysis. RFA, radiofrequency ablation; MWA, microwave ablation; LTP, local tumor progression; Loc-PFS, local progression-free survival; CRLM, colorectal cancer liver metastasis.

Table 4. LTP development and Loc-PFS of HCC lesions

	LTP Development			Loc-PFS				
	Developed	Not developed	P value	1-year survival	3-year survival	5-year survival	Median survival (months)	P value *
Ablation type								
MWA	6 (12.24%)	43 (87.76%)	0.771	57.14% (n = 28)	57.14% (n = 28)	57.14% (n = 28)	60.36	0.699
RFA	10 (14.08%)	61 (85.92%)		76.05% (n = 54)	54.92% (n = 39)	54.92% (n = 39)	77.33	
Lesion diameter								
<30 mm	7 (7.14%)	91 (92.86%)	<0.001	71.42% (n = 70)	64.28% (n = 63)	64.28% (n = 63)	26.72	<0.001
30–50 mm	9 (40.90%)	13 (59.10%)		54.54% (n = 12)	18.18% (n = 4)	18.18% (n = 4)	100.60	
The shortest vascular distance								
≤3 mm	10 (50.00%)	10 (50.00%)	<0.001	40.00% (n = 8)	25.00% (n = 5)	25.00% (n = 5)	45.82	0.014
>3 mm	6 (6.00%)	94 (94.00%)		74.00% (n = 74)	62.00% (n = 62)	62.00% (n = 62)	82.94	
Serum AFP level								
≤10 ng/mL	1 (2.32%)	42 (97.68%)	0.008	95.34% (n = 41)	95.34% (n = 41)	95.34% (n = 41)	118.06	<0.001
>10 ng/mL	15 (19.48%)	62 (80.52%)		53.24% (n = 41)	33.76% (n = 26)	33.76% (n = 26)	22.72	
Child–Pugh score								
Child-A	5 (5.81%)	81 (94.19%)	<0.001	82.55% (n = 71)	73.25% (n = 63)	73.25% (n = 63)	98.36	<0.001
Child-B	11 (32.35%)	23 (67.65%)		32.35% (n = 11)	11.76% (n = 4)	11.76% (n = 4)	12.02	
Cellular differentiation								
Well	2 (2.35%)	83 (97.65%)	<0.001	84.70% (n = 72)	78.82% (n = 67)	78.82% (n = 67)	106.33	<0.001
Moderate	14 (40.00%)	21 (60.00%)		28.57% (n = 10)	0.00% (n = 0)	0.00% (n = 0)	11.63	
Predisposing factor								
HBV	12 (23.07%)	40 (76.93%)	0.027	50.00% (n = 26)	42.30% (n = 22)	42.30% (n = 22)	49.10	0.210
HCV	3 (8.10%)	34 (91.90%)		78.37% (n = 29)	48.64% (n = 18)	48.64% (n = 18)	82.35	
NASH	1 (3.22%)	30 (96.78%)		87.09% (n = 27)	87.09% (n = 27)	87.09% (n = 27)	49.98	

*These P values indicate the results of univariate Cox regression analysis. HBV, hepatitis B virus; HCV, hepatitis C virus; NASH, non-alcoholic steatohepatitis; LTP, local tumor progression; Loc-PFS, local progression-free survival; HCC, hepatocellular carcinoma; MWA, microwave ablation.

ed their studies considering "patient-based" survival and observed better outcomes in patients with left-sided primary origin. In this "lesion-based" study, where more variables were considered, LTP was observed more frequently in CRLMs originating from the right colon. However, the study failed to achieve significant results. Concomitant lung metastasis, another CRC-specific variable, is one of the most important variables affecting survival and LTP.[26,27] In the study by Shady et al.[26], which only included patients with RFA, the presence of lung metastases was targeted as one of the most important prognostic factors. In this study, which included more comprehensive variables, the presence of lung metastases had an impact on LTP and Loc-PFS, and it proved to be more important than the primary origin of metastases and K-ras mutations. This suggests that concomitant lung metastases may be an important overall indicator of aggressive neoplastic behavior. Moreover, in the multiples analysis, it was found to be the second most important factor in lowering Loc-PFS after the shortest vascular distance of ≤3 mm.

There are numerous articles in the literature that include HCC-specific variants. One of the largest prospective studies that included patients with RFA and MWA, by Chong et al.[28] and Vietti Violi et al.[29], examined predisposing factors, the Child–Pugh score, and AFP levels, but they were not included in the statistical analysis. In another study comparing RFA with liver resection, in which 109 patients were treated with RFA, these three HCC-specific variables were included, and no effect of these three variables on "disease-free survival" was reported.[30] In a study examining 48 lesions with RFA, in addition to these three HCC-specific variables, the degree of histopathological differentiation was also included, of which only a high AFP level before ablation was correlated with "intrahepatic distant recurrence".[31] There are other studies that correlate with higher AFP levels.[32,33] In this study, a correlation was found between all these four variables and LTP. In addition, the multiples analysis revealed that the AFP level was the most important variable affecting the poor Loc-PFS outcome. This result is valuable in that it indicates that a host factor such as the AFP level is an important poor prognostic factor that outperforms even a tumor-based variable such as

the shortest vascular distance of ≤3 mm or large tumor size.

Complications were also investigated in this study as ancillary findings. Previous retrospective studies have shown that there was no difference in safety between ablation types.[8,34,35] A significant association was found between the occurrence of complications and intraoperative ablation, either when only the abscess or all complications were included. This could be due to a more invasive procedure and greater surface area of the peritoneum. There was also a strong correlation between the dilation of the bile duct and the ablation of the central segments (segments 1, 4b, and 5). It is understandable that the ablation of zones closer to the portal hilum may lead to this biliary obstruction. Costochondritis was observed in only one patient with a subcapsular lesion. The ablation of a subcapsular lesion in close proximity to the costochondral arcus may have caused this inflammation. This is the first time such a case was reported with the corresponding images.

This study has some limitations. First, it is a single-center, retrospective study. However, it represents the results of a large tertiary oncology center with a long-established thermal ablation protocol. Second, although an inspection was carried out, no patients with CRC with microsatellite instability and no patients with HCC with poor histopathological morphology were detected. Third, there were only three predisposing factors for chronic liver disease, and no other chronic liver diseases were included. However, this study provides a suitable basis for future thermal ablation studies to include more tumor-specific variables.

In conclusion, large tumor size and the shortest vascular distance of ≤3 mm are important factors with effects on LTP. However, host variables such as concomitant lung metastasis in patients with CRC and high pre-ablation AFP levels in patients with HCC may be important indicators of poor prognosis. Prospective randomized studies with tumor-specific variables and spatial characteristics are needed to explain the exact effects.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Van Cutsem E, Cervantes A, Adam R, et al. ESMO consensus guidelines for the management of patients with metastatic colorectal cancer. *Ann Oncol*. 2016;27(8):1386-422. [CrossRef]

2. Liver EAFTSOT. EASL clinical practice guidelines: management of hepatocellular carcinoma. *J Hepatol*. 2018;69(1):182-236. [CrossRef]

3. Minagawa M, Makuuchi M, Torzilli G, et al. Extension of the frontiers of surgical indications in the treatment of liver metastases from colorectal cancer: long-term results. *Ann Surg*. 2000;231(4):487-499. [CrossRef]

4. Ahmed M, Solbiati L, Brace CL, et al. Image-guided tumor ablation: standardization of terminology and reporting criteria--a 10-year update. *J Vasc Interv Radiol*. 2014;25:1691-1705. [CrossRef]

5. Kim C. Understanding the nuances of microwave ablation for more accurate post-treatment assessment. *Future Oncol*. 2018;14:1755-1764. [CrossRef]

6. Chinnaratha MA, Chuang MyA, Fraser RJ, Woodman RJ, Wigg AJ. Percutaneous thermal ablation for primary hepatocellular carcinoma: a systematic review and meta-analysis. *J Gastroenterol Hepatol*. 2016;31(2):294-301. [CrossRef]

7. Shady W, Petre EN, Do KG, et al. Percutaneous microwave versus radiofrequency ablation of colorectal liver metastases: ablation with clear margins (A0) provides the best local tumor control. *J Vasc Interv Radiol*. 2018;29(2):268-275. [CrossRef]

8. Poulou LS, Botsa E, Thanou I, Ziakas PD, Thanos L. Percutaneous microwave ablation vs radiofrequency ablation in the treatment of hepatocellular carcinoma. *World J Hepatol*. 2015;7(8):1054-1063. [CrossRef]

9. Facciorusso A, Di Maso M, Muscatiello N. Microwave ablation versus radiofrequency ablation for the treatment of hepatocellular carcinoma: a systematic review and meta-analysis. *Int J Hyperthermia*. 2016;32(3):339-344. [CrossRef]

10. Pathak S, Jones R, Tang JM, et al. Ablative therapies for colorectal liver metastases: a systematic review. *Colorectal Disease*. 2011;13(9):252-265. [CrossRef]

11. Berber E, Siperstein A. Local recurrence after laparoscopic radiofrequency ablation of liver tumors: an analysis of 1032 tumors. *Ann Surg Oncol*. 2008;15(10):2757-2764. [CrossRef]

12. Ahmed M, Solbiati L, Brace CL, et al. Image-guided tumor ablation: standardization of

terminology and reporting criteria -- a 10-year update. *Radiology*. 2014;273(1):241-260. [CrossRef]

13. Puijk RS, Ahmed M, Adam A, et al. Consensus guidelines for the definition of time-to-event end points in image-guided tumor ablation: Results of the SIO and DATECAN initiative. *Radiology*. 2021;301(3):533-540. [CrossRef]

14. Crocetti L, De Baere T, Lencioni R. Quality improvement guidelines for radiofrequency ablation of liver tumours. *Cardiovasc Interv Radiol*. 2010;33(1):11-17. [CrossRef]

15. Hosmer Jr DW, Lemeshow S, Sturdivant RX. Applied logistic regression, 3rd ed. John Wiley & Sons; 2013. [CrossRef]

16. Glassberg MB, Ghosh S, Clymer JW, et al. Microwave ablation compared with radiofrequency ablation for treatment of hepatocellular carcinoma and liver metastases: a systematic review and meta-analysis. *OncoTargets Ther*. 2019;12:6407-6438. [CrossRef]

17. Luo W, Zhang Y, He G, et al. Effects of radiofrequency ablation versus other ablating techniques on hepatocellular carcinomas: a systematic review and meta-analysis. *World J Surg Oncol*. 2017;15(1):126. [CrossRef]

18. Veltri A, Gazzera C, Calandri M, et al. Percutaneous treatment of Hepatocellular carcinoma exceeding 3 cm: combined therapy or microwave ablation? Preliminary results. *Radiol Med*. 2015;120(12):1177-1183. [CrossRef]

19. Di Vece F, Tombesi P, Ermili F, Maraldi C, Sartori S. Coagulation areas produced by cool-tip radiofrequency ablation and microwave ablation using a device to decrease back-heating effects: a prospective pilot study. *Cardiovasc Intervent Radiol*. 2014;37(3):723-729. [CrossRef]

20. Geyik S, Akhan O, Abbasoglu O, et al. Radiofrequency ablation of unresectable hepatic tumors. *Diagn Interv Radiol*. 2006;12(4):195-200. [CrossRef]

21. Calandri M, Odisio BC. Tailoring ablation strategies for colorectal liver metastases based upon rat sarcoma viral oncogene mutation status. *Chin Clin Oncol*. 2019;8(5):51. [CrossRef]

22. Rhaiem R, Rached L, Tashkandi A, Bouché O, Kianmanesh R. Implications of RAS Mutations on Oncological Outcomes of Surgical Resection and Thermal Ablation Techniques in the Treatment of Colorectal Liver Metastases. *Cancers*. 2022;14(3):816. [CrossRef]

23. Jiang BB, Yan K, Zhang ZY, et al. The value of KRAS gene status in predicting local tumor progression of colorectal liver metastases following radiofrequency ablation. *Int J Hyperthermia*. 2019;36(1):211-219. [CrossRef]

24. Zhou F, Yu X, Liang P, et al. Does primary tumor location impact the prognosis of colorectal liver metastases patients after microwave ablation?-Lessons from 10 years' experience. *Oncotarget*. 2017;8:100791-100800. [CrossRef]

25. Gu Y, Huang Z, Gu H, Gao F, et al. Does the site of the primary affect outcomes when ablating colorectal liver metastases with radiofrequency ablation? *Cardiovasc Intervent Radiol*. 2018;41(6):912-919. [CrossRef]

26. Shady W, Petre EN, Gonen M, et al. Percutaneous radiofrequency ablation of colorectal cancer liver metastases: factors affecting outcomes -- a 10-year experience at a single center. *Radiology*. 2016;278:601-611. [CrossRef]

27. Merkel S, Weber K, Croner R, et al. Distant metastases in colorectal carcinoma: a proposal for a new M1 subclassification. *Eur J Surg Oncol*. 2016;42:1337-1342. [CrossRef]

28. Chong CCN, Lee KF, Cheung SYS, et al. Prospective double-blinded randomized controlled trial of Microwave versus RadioFrequency Ablation for hepatocellular carcinoma (McRFA trial). *HPB (Oxford)*. 2020;22(8):1121-1127. [CrossRef]

29. Vietti Violi N, Duran R, Guiu B, et al. Efficacy of microwave ablation versus radiofrequency ablation for the treatment of hepatocellular carcinoma in patients with chronic liver disease: a randomised controlled phase 2 trial. *Lancet Gastroenterol Hepatol*. 2018;3(5):317-325. [CrossRef]

30. Ng K, Chok K, Chan A, et al. Randomized clinical trial of hepatic resection versus radiofrequency ablation for early-stage hepatocellular carcinoma. *J Br Surg*. 2017;104(13):1775-1784. [CrossRef]

31. Zytoon AA, Ishii H, Murakami K, et al. Recurrence-free survival after radiofrequency ablation of hepatocellular carcinoma. A registry report of the impact of risk factors on outcome. *Japan J Clin Oncol*. 2007;37(9):658-672. [CrossRef]

32. Izumi N, Asahina Y, Noguchi O, Uchihara M, et al. Risk factors for distant recurrence of hepatocellular carcinoma in the liver after complete coagulation by microwave or radiofrequency ablation. *Cancer*. 2001;91(5):949-956. [CrossRef]

33. Harrison LE, Koneru B, Baramipour P, et al. Locoregional recurrences are frequent after radiofrequency ablation for hepatocellular carcinoma. *J Am Coll Surg*. 2003;197(5):759-764. [CrossRef]

34. Lahat E, Eshkenazy R, Zendel A, et al. Complications after percutaneous ablation of liver tumors: a systematic review. *Hepatobiliary Surg Nutr*. 2014;3(5):317-323. [CrossRef]

35. Hauser K, Matthes J. Medical students' medication communication skills regarding drug prescription - qualitative analysis of simulated physician-patient consultations. *Eur J Clin Pharmacol*. 2017;73(4):429-435. [CrossRef]

Supplementary Table 1. Results of the Cox regression and multiples analysis of the variables' effect on Loc-PFS*

	Hazard ratio	Confidence interval (95%)	P value
CRLM			
Ablation technique (MWA vs. RFA)	0.619	0.166–2.311	0.476
K-ras (Mutant vs. wild)	1.940	0.520–7.229	0.323
Lesion diameter (30–50 mm vs. <30 mm)	1.065	0.348–3.259	0.911
Concomitant lung metastasis (yes vs. no)	6.437	1.661–24.946	0.027
SVD (≤3 mm vs. >3 mm)	6.604	1.065–40.941	0.007
HCC			
Serum AFP level (>10 ng/mL vs. ≤10 ng/mL)	6.323	1.038–38.494	0.045
Cellular differentiation (moderate vs. poor)	2.134	0.915–8.191	0.068
Child–Pugh score (Child-B vs. Child-A)	2.510	0.611–10.303	0.101
Lesion diameter (30–50 mm vs. <30 mm)	1.329	0.433–4.079	0.618
SVD (≤3 mm vs. >3 mm)	1.924	0.894–3.917	0.090

*Only variables with *P* < 0.20 values were included in the multiples analysis. AFP, alpha-feto protein; CRLM, colorectal cancer liver metastasis; HCC, hepatocellular carcinoma; SVD, shortest vascular distance; RFA, radiofrequency ablation; MWA, microwave ablation; Loc-PFS, local progression-free survival.

Predictability of the radiological response to Yttrium-90 transarterial radioembolization by dynamic magnetic resonance imaging-based radiomics analysis in patients with intrahepatic cholangiocarcinoma

 Hüseyin Tuğsan Ballı[1]
 Ferhat Can Pişkin[1]
 Sevinç Püren Yücel[2]
 Sinan Sözütok[1]
 Duygu Özgül[1]
 Kairgeldy Aikimbaev[1]

[1]Çukurova University Faculty of Medicine, Balçalı Hospital Health Application and Research Center, Department of Radiology, Adana, Türkiye

[2]Çukurova University Faculty of Medicine, Balçalı Hospital Health Application and Research Center, Department of Biostatistics, Adana, Türkiye

PURPOSE

The study aims to investigate the predictability of the radiological response in intrahepatic cholangiocarcinoma (iCC) patients undergoing Yttrium-90 transarterial radioembolization (TARE) with a combined model built on dynamic magnetic resonance imaging (MRI)-based radiomics and clinical features.

METHODS

Thirty-six naive iCC patients who underwent TARE were included in this study. The tumor segmentation was performed on the axial T2-weighted (T2W) without fat suppression, axial T2W with fat suppression, and axial T1-weighted (T1W) contrast-enhanced (CE) sequence in equilibrium phase (Eq). At the sixth month MRI follow-up, all patients were divided into responders and non-responders according to the modified Response Evaluation Criteria in Solid Tumors. Subsequently, a radiomics score (rad-score) and a combined model of the rad-score and clinical features for each sequence were generated and compared between the groups.

RESULTS

Thirteen (36.1%) patients were considered responders, and the remaining 23 (63.9%) were non-responders. Responders exhibited significantly lower rad-scores than non-responders ($P < 0.050$ for all sequences). The radiomics models showed good discriminatory ability with an area under the curve (AUC) of 0.696 [95% confidence interval (CI), 0.522–0.870] for the axial T1W-CE-Eq, AUC of 0.839 (95% CI, 0.709–0.970) for the axial T2W with fat suppression, and AUC of 0.836 (95% CI, 0.678–0.995) for the axial T2W without fat suppression.

CONCLUSION

Radiomics models created by pre-treatment MRIs can predict the radiological response to Yttrium-90 TARE in iCC patients with high accuracy. Combining radiomics with clinical features could increase the power of the test. Large-scale studies of multi-parametric MRIs with internal and external validations are needed to determine the clinical value of radiomics in iCC patients.

KEYWORDS

Radiomics, magnetic resonance imaging, radioembolization, intrahepatic cholangiocarcinoma, radiological response

Corresponding author: Ferhat Can Pişkin

E-mail: ferhatcpiskin@gmail.com

Intrahepatic cholangiocarcinoma (iCC) is the second most common primary hepatic malignancy after hepatocellular carcinoma.[1] Its worldwide incidence has increased over the past few decades.[2] If left untreated, the prognosis is poor, with an estimated median survival of 3 to 8 months. Treatment options for iCC include surgical resection and transplantation. Unfortunately, most patients will present with metastatic or locally advanced disease at diagnosis and are not candidates for surgery.[3] For unresectable iCCs, systemic chemotherapy with cisplatin-gemcitabine results in a relatively poor median overall survival (OS) of 11.7 months.[4]

Currently, transarterial radioembolization (TARE) is used as the first-line treatment due to radiation sensitivity and high arterial perfusion of the tumor.[5] The results of TARE are mixed, with median response rates ranging from 5% to 36% and median OS from 9 to 22 months.[6] Additionally, TARE is a costly and laborious treatment method; therefore, predicting response to treatment is crucial for accurate patient selection.

Radiomics is the post-processing analysis of medical images with custom-made software to obtain texture data imperceptible to the human eye. The data obtained are analyzed with machine learning algorithms and developed models.[7] The number of studies in radiomics, particularly for predicting the treatment response of hepatic malignancies, including hepatocellular carcinoma and hepatic metastasis, has increased exponentially in recent years.[8,9] However, although iCC is the second most common primary liver cancer, they are relatively rare tumors, and studies on radiomics in iCC patients are limited and derived from computed tomography (CT) examinations.[10,11] As far as we know, there are yet to be studies on whether radiomics analyses based on magnetic resonance imaging (MRI) can predict the radiological response to TARE in iCC patients.

This study aims to investigate the predictability of the treatment response in iCC patients undergoing Yttrium-90 TARE with a combined model created with dynamic MRI-based radiomics and clinical features.

Methods

Study design

The Institutional Clinical Research Çukurova University, Faculty of Medicine, Clinical Ethics Commitee (decision number: 114/09-2021) approved this single-center retrospective study. Informed consent was obtained from all patients prior to all diagnostic and therapeutic procedures in accordance with the principles of the 1964 Declaration of Helsinki.

Fifty-five naive iCC patients who underwent TARE between September 2015 and January 2022 were included in the study. The inclusion criteria were a biopsy-proven diagnosis of iCC and dynamic MRI before and after TARE. The exclusion criteria were prior local or systemic treatments, an inability to clearly distinguish tumor boundaries due to the infiltrative pattern on the pre-treatment MRI, and images unsuitable for analysis due to motion artifacts. Nineteen patients who underwent TARE for iCC were excluded from the study after application of the exclusion criteria. As a result, a total of 36 patients who met the selected criteria were included in the study.

Pre-treatment clinical characteristics, including age, gender, alpha-fetoprotein, carcinoembryonic antigen, carbohydrate antigen 19-9, alanine aminotransferase, aspartate aminotransferase, total bilirubin, and albumin, an international normalized ratio (INR), intrahepatic tumor distribution, positron emission tomography/CT-based extrahepatic disease spread, and nodal involvement were noted. The laboratory examination results were obtained from blood tests the day before TARE and during planned follow-ups.

MRI examinations

The MRI examinations were acquired using a 1.5 Tesla system (Optima, General Electric Healthcare, USA) or a 3.0 Tesla system (Ingenia, Philips Medical Systems, the Netherlands). The MRI sequences were composed of an axial T2-weighted (T2W) without fat suppression, axial T2W with fat suppression, and axial T1-weighted (T1W) contrast-enhanced (CE) sequence in equilibrium phases (Eq). The specific parameters of axial T2W imaging were as follows: time of repetition (TR) 10,000 ms, time of echo (TE) 66 ms, layer thickness 6 mm, layer spacing 1 mm, matrix 320 × 320, field of view (FOV) 400 mm × 400 mm, piecewise collection times or average times 1, and parallel collection factor 0, fs. The parameters of dynamic CE MRI were as follows: TR 4.2 ms, TE 1 min full, layer thickness 5 mm, layer spacing 0 mm, matrix 260 × 224 mm, FOV 380 mm × 342 mm, and parallel acceleration factor 2. T1W was acquired using 0.1 mmol/kg gadolinium-diethylenetriamine penta-acetic acid (Gd-DTPA) at a rate of 2.5 mL/s in the Eq (a scanning delay of 180 s). MRI sequences have been abbreviated as "phase 1: axial T1W CE Eq, phase 2: axial T2W with fat suppression, and phase 3: axial T2W sequence without fat suppression" in relevant places in the text.

Transarterial radioembolization

All patients underwent splanchnic angiography via the femoral approach, and the tumor-feeding arteries were determined by cone-beam CT, followed by a 99m technetium-macroaggregate albumin (MAA) injection. The lung shunt fraction and distribution of MAA within the tumors and non-tumor tissue were evaluated with single-photon emission CT. The desired dose was calculated using partition model dosimetry.[12] During TARE, infusion of a previously calculated dose of the Yittrum-90-loaded resin (SIR-Spheres, Sirtex Medical, Australia) or glass microspheres (TheraSphere, Boston Scientific, US) was carried out under fluoroscopic guidance with super-selective or selective manner depending on the defined vascular anatomy. All patients were scheduled for follow-up, including MRI and laboratory tests. After the TARE procedure, the patients were observed for complications for 24 hours.

Evaluation of the radiological response to treatment

Following TARE, dynamic CE MRI was performed at intervals of three consecutive months. The response of the index tumor to the treatment was evaluated according to the modified Response Evaluation Criteria in Solid Tumors.[13] The objective response of the index tumor represented the primary outcome measure and was defined as the sum of the complete response and partial response. Based on the 6-month MRI follow-up, the patients were divided into two groups of responders and non-responders.

Tumor segmentation

Digital Imaging and Communications in Medicine data were transferred to a workstation and analyzed by dedicated software (Olea Sphere v.3 SP2, Olea Medical, France). The raw images were normalized using a Z-score to rule out the possible effects of different MRI devices. Subsequently, axial T2W without fat suppression, axial T2W with fat suppression, and axial T1W-CE-Eq images were segmented by two radiologists blinded to the aim of this study manually drawing the boundaries of the tumors slice-by-slice. After this, a volume of interest (VOI) that covered the entire tumor was created (Figure 1). One

---

Main points

- Intrahepatic cholangiocarcinoma (iCC) is the second most common primary hepatic malignancy.

- Transarterial radioembolization (TARE) is used as a first-line treatment in iCC patients due to the radiation-sensitivity of this tumor.

- TARE is a costly and laborious treatment method; therefore, predicting the response to the treatment is crucial for accurate patient selection.

- In radiomics models created by pre-treatment magnetic resonance imaging, the response to TARE in iCC patients can be predicted with high accuracy.
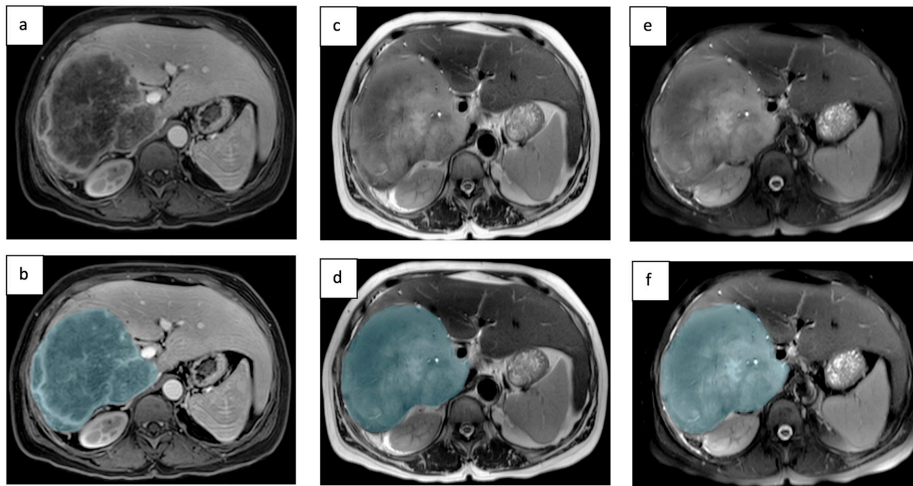
Figure 1. A 58-year-old male patient had pathologically proven intrahepatic cholangiocarcinoma in the entire right lobe of the liver. Segmentation of this mass was performed on axial T1-weighted contrast-enhanced equilibrium phase **(a, b)**, axial T2-weighted sequence without fat suppression **(c, d)**, and axial T2-weighted with fat suppression **(e, f)** magnetic resonance images.

hundred eight grey-level properties (first and second order) of the generated VOI were extracted.

Selection of the treatment response-related features and construction of a radiomics score

One hundred eight features were extracted based on the MRI for each patient. Because the number of features was superior to the number of patients, a radiomic feature selection process was constructed using the lowest absolute shrinkage and selection operator (LASSO)[14] The logistic radiomics models for predicting the treatment response for all phases were fitted to select the treatment response-related features with nonzero coefficients. Three-fold cross-validation with minimum criteria was employed to find an optimal tunning parameter, where the final value of the tuning parameter yielded minimum cross-validation error and maximum area under the curve (AUC). Then, the radiomics score (rad-score) was calculated for each patient by a linear combination of the selected features (with nonzero coefficients) and their respective coefficients.

Statistical analysis

All analyses were performed using IBM SPSS software (version 20; IBM Corp, USA) and R software (version 1.0.143). Categorical variables were expressed as numbers and percentages, whereas continuous variables were summarized as mean, standard deviation, median, and minimum-maximum where appropriate. The chi-squared test was used to compare categorical variables between patient groups. The normality of

distribution for continuous variables was confirmed with the Shapiro–Wilk test. The Student's t-test or Mann–Whitney U test was used to compare the continuous clinical characteristics between patient groups depending on whether the statistical hypotheses were fulfilled. The glmnet package (https://cran.r-project.org/web/packages/glmnet/index.html) was used for the LASSO binary logistic regression. The distribution of the rad-scores in the treatment response groups was demonstrated via a violin plot, which is a hybrid of a box plot and a kernel density plot. Violin plots were plotted using the ggplot2 package (https://cran.r-project.org/web/packages/ggplot2/index.html). Logistic regression analysis was performed to determine significant predictors of the treatment response. Clinical features that were significant at the $P < 0.250$ level in the univariate analysis were entered into the stepwise logistic regression analysis using the backward logistic regression method. Features with a $P < 0.050$ after the stepwise analysis were included in the clinical model. In addition, three combined models were built: (1) a model adding the rad-score in the axial T2W without fat suppression to the clinical model, (2) a model adding the rad-score in the axial T2W with fat suppression to the clinical model, and (3) a model adding the rad-score in the axial T1W CE Eq to the clinical model. The goodness-of-fit of the models was assessed with Nagelkerke's R-squared model.

The predictive ability of the models was assessed with receiver operator characteristic curves and associated performance diagnostics (AUC, sensitivity, and specificity). The best cut-off value was based on the index

of union method.[15] The AUCs of the models were compared with the DeLong test (https://www.rdocumentation.org/packages/Daim/versions/1.1.0/topics/DeLong.test). The net reclassification index (NRI) and integrated discrimination improvement (IDI) were used to assess the discrimination and reclassification ability to use the rad-score.[16] Each combined model was compared with the clinical model as a reference to assess them. The PredictABEL package was used to calculate the NRI and IDI (https://cran.r-project.org/web/packages/PredictABEL/index.html). The statistical level of significance for all tests was $P < 0.050$.

Results

Clinical characteristics

Thirteen (36.1%) patients were considered responders, and the remaining 23 (63.9%) were non-responders at the 6-month follow-up. Table 1 presents the baseline clinical characteristics of the patients in the treatment groups. There were no significant differences in any of the characteristics between the two treatment response groups ($P > 0.050$ for all).

Radiomics signature calculation and evaluation

To investigate the effectiveness of the treatment response discrimination, we performed LASSO modeling of the texture features; 108 features were chosen to construct the rad-score for the axial T1W-CE-Eq. Similarly, four features were selected for the axial T2W with fat suppression and eight features for the axial T2W without fat suppression. Using these features, rad-scores were generated for each patient in three phases, and Supplementary Material 1 contains the details of the feature selection process.

Responders exhibited significantly lower rad-scores than non-responders in all phases ($P = 0.039$ for the axial T1W-CE-Eq, $P = 0.001$ for the axial T2W with fat suppression, and $P = 0.001$ for the axial T2W without fat suppression). Figure 2 presents the violin plot of the rad-scores for all phases.

Model building and performances

Table 2 summarizes the results of the multivariate logistic regression analysis. After the stepwise regression analysis, results for the clinical model (before the rad-score was added to the clinical features) revealed that bilobar disease [odds ratio (OR): 4.53, 95% confidence interval (CI): 1.06–19.41, $P = 0.042$]

Table 1. Clinical characteristics of the study population

	Treatment response		P
	Responders (n = 13)	Non-responders (n = 23)	
Age, years	55.0 ± 15.6	59.6 ± 9.3	0.323
Gender, n (%)			
Male	8 (66.7)	4 (33.3)	> 0.999
Female	15 (62.5)	9 (37.5)	
Bilobar disease, n (%)			
-	17 (77.3)	5 (22.7)	0.082
+	6 (42.9)	8 (57.1)	
Extrahepatic disease, n (%)			
-	16 (69.6)	7 (30.4)	0.474
+	7 (53.8)	6 (46.2)	
Lymphadenopathy, n (%)			0.474
-	16 (69.6)	7 (30.4)	
+	7 (53.8)	6 (46.2)	
CEA	2.6 (0.5–69.1)	1.9 (0.6–106.4)	0.675
CEA, n (%)			
Normal	8 (33.3)	16 (66.7)	0.720
Abnormal	5 (41.7)	7 (58.3)	
CA 19.9	141.4 (2.0–1786.8)	31.1 (0.0–1859.1)	0.093
CA 19.9, n (%)			
Normal	3 (20.0)	12 (80.0)	0.177
Abnormal	10 (47.6)	11 (52.4)	
CA19-9log10	2.1 ± 0.8	1.6 ± 0.9	0.147
AFP	4.9 (2.0–198718.0)	4.0 (1.0–437.7)	0.344
AFP, n (%)			
Normal	7 (30.4)	16 (69.6)	0.474
Abnormal	6 (46.2)	7 (53.8)	
AFPlog10	0.7 (0.3–5.3)	0.6 (0.0–2.6)	0.344
Total bilirubin	0.6 (0.3–0.9)	0.5 (0.1–3.2)	0.267
Total bilirubin, n (%)			
Normal	11 (42.3)	15 (57.7)	0.270
Abnormal	2 (20.0)	8 (80.0)	
Albumin	38.9 (31.1–45.0)	40.8 (25.0–44.3)	0.190
INR	1.1 ± 0.1	1.0 ± 0.1	0.103
Rad-score (phase 1)	-0.5415 ± 0.0466	-0.5879 ± 0.0691	0.039
Rad-score (phase 2)	-0.2136 ± 0.3581	-0.8405 ± 0.5425	0.001
Rad-score (phase 3)	-0.1994 ± 0.5798	-0.8491 ± 0.4408	0.001

Unless otherwise specified, data were expressed as mean ± standard deviation or median (min-max). CEA, carcinoembroyologic antigen; CA, carbohydrate antigen; AFP, alfa-fetoprotein; INR, international normalized ratio; rad score, radiomics score; phase 1, axial T1-weighted contrast-enhanced equilibrium phase; phase 2, axial T2-weighted with fat suppression; phase 3, axial T2-weighted sequence without fat suppression.

and the INR (OR: 2.31, 95% CI 0.93–5.74, P = 0.072) were significant independent risk factors for the treatment response. Results of the combined models (obtained by integrating the significant clinical features and the rad-score in each phase) demonstrated that bilobar disease and the rad-score in the axial T2W with fat suppression (OR: 7.97, 95% CI: 1.03–62.03, P = 0.047 and OR: 1.33, 95% CI: 1.06–1.68, P = 0.015) and the rad-score in the axial T2W without fat suppression (OR: 1.31, 95% CI: 1.04–1.65, P = 0.023) were independent predictors of the treatment response.

The radiomics models (fitted only from the rad-scores in each phase) showed good discriminatory ability with an AUC of 0.696 (95% CI: 0.522–0.870) for the axial T1W-CE-Eq, 0.839 (95% CI: 0.709–0.970) for the axial T2W with fat suppression, and 0.836 (95% CI, 0.678–0.995) for the axial T2W without fat suppression (Table 3). There was no significant difference in AUCs between the radiomics models (DeLong's tests P > 0.050 for all pairwise comparisons).

The clinical model resulted in an AUC of 0.769, followed by the combined model-1 (0.816), the combined model-2 (0.863), and the combined model-3 (0.880) (Figure 3). Although the AUC of the combined model-3 was not significantly higher than the other models, the combined model-3 showed a favorable AUC of 0.880 (95% CI: 0.730–0.999) (Table 3). The sensitivity and specificity of the combined model-3 were 92% and 78%, respectively. Relative to the clinical model, the use of the combined model-2 resulted in an NRI of 93.0% (P = 0.002) and an IDI of 20.0% (P = 0.003), and the use of combined model-3 resulted in an NRI of 86.0% (P = 0.006) and an IDI of 22.0% (P < 0.001). The reclassification measures of discrimination confirmed that adding rad-scores to the clinical model (the combined model-2 and the combined model-3) performed better than the clinical model alone. Table 3 presents the detailed information for the prediction performance of the models.

Discussion

In this study, the predictability of the treatment responses in iCC patients undergoing TARE was investigated with a combined model created with MRI-based (including the axial T1W-CE-Eq, axial T2W without fat suppression, and axial T2 with fat suppression sequences) radiomics and clinical features. Radiomics models were produced to predict the radiological response with high accuracy. Bilobar disease and rad-scores were independent predictors of the treatment response. There was no statistical difference between the models combining clinical characteristics with radiomics features. This study is important because it is the first one in which the response of TARE in iCC patients has been predicted with MRI-based radiomics.

Patients with unresectable iCC have a poor prognosis. The previously published studies revealed that TARE has great potential to improve patients' prognosis and OS. However, they have reported a wide range of median OS in iCC (6.1–22 months), probably reflecting the heterogeneous biological behavior of this relatively rare tumor.[17-19] There-

fore, the pre-treatment determination of the prognostic factors is important in the patient selection for TARE and the implementation of personalized treatments. In this study, the tumor responded to therapy in about a third of the treated patients. Texture analysis based on pre-treatment MRI was a valuable marker for predicting the treatment response in unresectable iCC patients who underwent TARE.

Previous studies have identified clinical prognostic factors in patients with iCC who underwent TARE. Tumors with bilobar disease had a lower OS rate after the administration of TARE than tumors with unilobar disease. On the other hand, it was established that extrahepatic disease and liver function did not affect the prognosis.[20] In this study, it was found that extrahepatic metastases

and liver function did not have prognostic significance. However, bilobar disease was associated with a treatment response in iCC patients.

Mosconi et al.[11] analyzed the data of 53 iCC patients who underwent TARE and investigated the relationship between CT textural features prior to TARE and the OR. They used the arterial phase images for texture analysis to show that iCCs with a high uptake of iodine contrast in the arterial phase had a higher OR after TARE. Combining these textural features provided an AUC for an OR prediction of 0.896 (95% CI 0.814–0.977). In the present study, MRI was used, as it has a better resolution than CT and shows more tumor tissue features. The AUC in the authors' study (0.880) was similar to Mosconi et al.'s[11] results.

Zhang et al.[21] investigated predicting the immunophenotyping (IP) and OS of iCC patients using preoperative MRI texture analysis. They found that the MRI tissue signature could serve as a potential predictive biomarker for IP and OS using arterial phase images for tissue analysis.[21] Mosconi et al.[11] considered that tumor enhancement at the arterial phase indicated hyperperfusion as the applicability of TARE. Zhang et al.[21] thought that the arterial phase revealed the amount of inflammation better than other MRI sequences. In the present study, tissue analysis was performed in the axial T2W with and without fat suppression and axial T1W CE-Eq because the amount of fibrous component associated with poor prognosis is better visualized on MRI as a peripheral hypointensity in T2W and CE images on the delay phase.[21] In this study, the arterial phase was not used since the truncation artifact negatively affects tumor segmentation in MRI with Gd-DTPA. Therefore, this study reveals the importance of using other MRI sequences (axial T2W without fat suppression
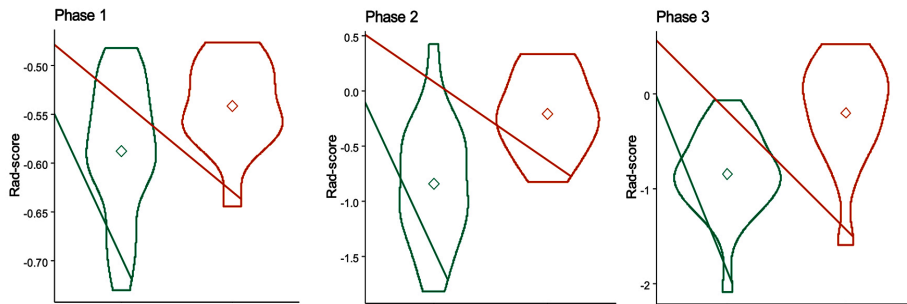

Figure 2. The violin plot of the rad-scores for non-responders and responders. The wider parts of the violin plot show that the patients of the group are more likely to receive the given value, while the thinner parts are less likely. The squares represent the mean values. The difference between radiomics scores was compared with the independent samples t-test ($P = 0.039$ for axial T1-weighted-contrast-enhanced equilibrium, $P = 0.001$ for axial T2-weighted with fat suppression, and $P = 0.001$ for the axial T2-weighted without fat suppression).

Table 2. Multivariate logistic regression analysis of the prediction of the treatment response

Variables	Clinical model		Combined model-1		Combined model-2		Combined model-3	
	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P	OR (95% CI)	P
Bilobar disease	4.53 (1.06–19.41)	0.042	4.76 (0.92–27.8)	0.064	7.97 (1.03–62.03)	0.047	5.15 (0.80–33.20)	0.085
INR	2.31 (0.93–5.74)	0.072	1.89 (0.72–5.1)	0.195	1.61 (0.53–4.88)	0.404	1.26 (0.38–4.15)	0.701
Rad-score - phase 1			2.50 (0.54–11.49)	0.239				
Rad-score - phase 2					1.33 (1.06–1.68)	0.015		
Rad-score - phase 3							1.31 (1.04–1.65)	0.023

OR, objective response; CI, confidence interval; INR, international normalized ratio; rad-score, radiomics score; phase 1, axial T1-weighted contrast-enhanced equilibrium phase; phase 2, axial T2-weighted with fat suppression; phase 3, axial T2-weighted without fat suppression.

Table 3. Performance of the radiomics models

	Cut-off	SEN	SPE	AUC (95% CI)	P	NRI (95% CI)	P	IDI (95% CI)	P
Radiomics models									
Phase 1	-0.5585	0.54	0.56	0.696 (0.522–0.870)	0.054				
Phase 2	-0.4024	0.77	0.83	0.839 (0.709–0.970)	0.001				
Phase 3	-0.4254	0.77	0.83	0.836 (0.678–0.995)	0.001				
Clinical model	0.4181	0.69	0.74	0.769 (0.607–0.931)	0.008				
Combined model-1	0.4052	0.85	0.74	0.816 (0.667–0.965)	0.002	0.19 (-0.48–0.86)	0.582	0.04 (-0.02–0.09)	0.227
Combined model-2	0.3375	0.85	0.78	0.863 (0.744–0.982)	<0.001	0.93 (0.34–1.52)	0.002	0.20 (0.06–0.33)	0.003
Combined model-3	0.3211	0.92	0.78	0.880 (0.730–0.999)	<0.001	0.86 (0.25–1.48)	0.006	0.22 (0.09–0.35)	<0.001

Note: (1) Nagelkerke R^2: phase 1 0.172, phase 2 0.400, phase 3 0.409, clinic model 0.273, combined model-1 0.319, combined model-2 0.520, combined model-3 0.491. (2) NRI and IDI values refer to the clinical model compared to the corresponding combined model. SEN, sensitivity; SPE, specificity; AUC, area under the curve; CI, confidence interval; NRI, net reclassification index; IDI, integrated discrimination improvement; phase 1, axial T1-weighted contrast-enhanced equilibrium phase; phase 2, axial T2-weighted with fat suppression; phase 3, axial T2-weighted without fat suppression.
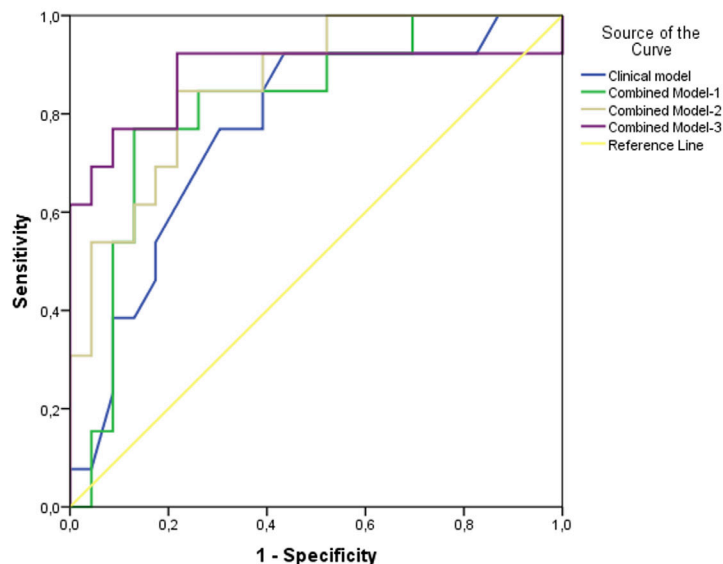
Figure 3. Receiver operator characteristic curves of models.

and axial T1W-CE-Eq) other than the arterial phase for texture analysis.

The rad-scores constructed with the LASSO were significantly associated with a treatment response for all phases in this study. Although the inclusion of the rad-score in the clinical model did not statistically substantially improve the AUC, it increased the sensitivity in predicting the treatment response and improved model performance. The combined clinical model-2 and model-3 showed enhanced AUCs of 0.863 and 0.880 with an explicit NRI and IDI.

There were several limitations to this study. First, the number of patients was limited due to the study's retrospective nature. Therefore, internal or external validation analysis could not be performed. Second, the images analyzed in the study were obtained from two devices with different Tesla powers. This could have affected the texture analysis. However, to avoid this, normalization was applied to all images before segmentation. Third, the study did not evaluate other MRI sequences and dynamic contrast phases (portal phase). Despite all these limitations, the present study demonstrated that the treatment outcomes of iCC patients undergoing TARE could be predicted with high accuracy by MRI-based radiomics prior to treatment.

In radiomics models created by pre-treatment MRIs, the response to TARE in iCC patients can be predicted with high accuracy. The combination of clinical factors, such as bilobar disease and texture analysis, could increase the power of the test. However, large-scale studies with multiparametric MRIs with internal and external validations are needed to reach a definitive conclusion and determine the advantages and disadvantages over the radiomics models.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Khan SA, Toledano MB, Taylor-Robinson SD. Epidemiology, risk factors, and pathogenesis of cholangiocarcinoma. *HPB (Oxford)*. 2008;10(2):77-82. [CrossRef]

2. Everhart JE, Ruhl CE. Burden of digestive diseases in the United States part III: liver, biliary tract, and pancreas. *Gastroenterology*. 2009;136(4):1134-1144. [CrossRef]

3. Ejaz A, Cloyd JM, Pawlik TM. Advances in the diagnosis and treatment of patients with intrahepatic cholangiocarcinoma. *Ann Surg Oncol*. 2020;27(2):552-560. [CrossRef]

4. Valle J, Wasan H, Palmer DH, et al. Cisplatin plus gemcitabine versus gemcitabine for biliary tract cancer. *N Engl J Med*. 2010;362(14):1273-1281. [CrossRef]

5. Al-Adra DP, Gill RS, Axford SJ, Shi X, Kneteman N, Liau SS. Treatment of unresectable intrahepatic cholangiocarcinoma with Yttrium radioembolizationtion: a systematic review and pooled analysis. *Eur J Surg Oncol*. 2015;41(1):120-127. [CrossRef]

6. Edeline J, Touchefeu Y, Guiu B, et al. Radioembolization plus chemotherapy for first-line treatment of locally advanced intrahepatic cholangiocarcinoma: a phase 2 clinical trial. *JAMA Oncol*. 2020;6(1):51-59. [CrossRef]

7. Liu Z, Wang S, Dong D, et al. The applications of radiomics in precision diagnosis and treatment of oncology: opportunities and challenges. *Theranostics*. 2019;12;9(5):1303-1322. [CrossRef]

8. Saini A, Breen I, Pershad Y, et al. Radiogenomics and radiomics in liver cancers. *Diagnostics (Basel)*. 2018;9(1):4. [CrossRef]

9. Lewis S, Hectors S, Taouli B. Radiomics of hepatocellular carcinoma. *Abdom Radiol (NY)*. 2021;46(1):111-123. [CrossRef]

10. Kobe A, Zgraggen J, Messmer F, et al. Prediction of treatment response to transarterial radioembolization of liver metastases: radiomics analysis of pre-treatment cone-beam CT: a proof of concept study. *Eur J Radiol Open*. 2021;8:100375. [CrossRef]

11. Mosconi C, Cucchetti A, Bruno A, et al. Radiomics of cholangiocarcinoma on pretreatment CT can identify patients who would best respond to radioembolisation. *Eur Radiol*. 2020;30(8):4534-4544. [CrossRef]

12. Kao YH, Tan EH, Ng CE, Goh SW. Clinical implications of the body surface area method versus partition model dosimetry for Yttrium radioembolization using resin microspheres: a technical review. *Ann Nucl Med*. 2011;25(7):455-461. [CrossRef]

13. Lencioni R, Llovet JM. Modified RECIST (mRECIST) assessment for hepatocellular carcinoma. *Semin Liver Dis*. 2010;30(1):52-60. [CrossRef]

14. Tibshirani R. Regression shrinkage and selection via the LASSO. *J R Stat Soc Ser B*. 1996;58(1):267-288. [CrossRef]

15. Unal I. Defining an optimal cut-point value in ROC analysis: an alternative approach. *Comput Math Methods Med*. 2017;2017:e3762651. [CrossRef]

16. Pencina MJ, D'Agostino RB Sr, D'Agostino RB Jr, Vasan RS. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Stat Med*. 2008;27(2):157-172. [CrossRef]

17. Mosconi C, Gramenzi A, Ascanio S, et al. Yttrium-90 radioembolization for unresectable/recurrent intrahepatic cholangiocarcinoma: a survival, efficacy and safety study. *Br J Cancer*. 2016;115(3):297-302. [CrossRef]

18. Mouli S, Memon K, Baker T, et al. Yttrium-90 radioembolization for intrahepatic cholangiocarcinoma: safety, response, and survival analysis. *J Vasc Interv Radiol*. 2013;24(8):1227-1234. [CrossRef]

19. Bargellini I, Mosconi C, Pizzi G, et al. Yttrium-90 radioembolization in unresectable intrahepatic cholangiocarcinoma: results of a multicenter retrospective study. *Cardiovasc Intervent Radiol*. 2020;43(9):1305-1314. [CrossRef]

20. Köhler M, Harders F, Lohöfer F, et al. Prognostic factors for overall survival in advanced intrahepatic cholangiocarcinoma treated with Yttrium-90 radioembolization. *J Clin Med*. 2019;25;9(1):56. [CrossRef]

21. Zhang J, Wu Z, Zhao J, et al. Intrahepatic cholangiocarcinoma: MRI texture signature as predictive biomarkers of immunophenotyping and survival. *Eur Radiol*. 2021;31(6):3661-3672. [CrossRef]

Supplementary Material 1

Selection process of the radiomics features and radiomics score calculation
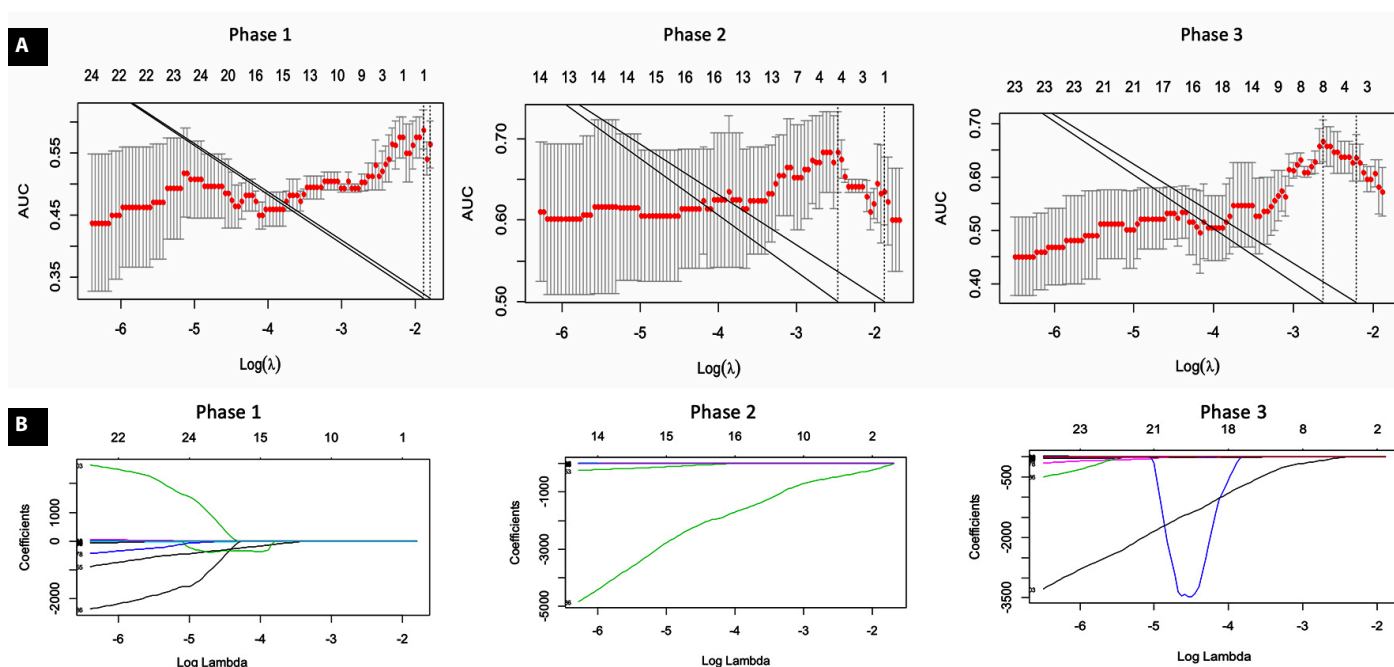
Figure 2 shows the selection process of radiomics texture features in phases by the lowest absolute shrinkage and selection operator logistic regression model. The tuning parameter determined by maximizing the area under the curve in Figure A was used to select features with non-zero coefficients from the coefficient profiles plot in Figure B. The radiomics scores of the patients were calculated in each phase by multiplying the selected features with their respective coefficients:

Radiomics score = intercept + coefficient × radiomics features

In each phase, rad-scores were calculated for each individual separately; that is, three rad-scores were obtained for each individual. Table 1 shows the details of the selected features.

Sonographic cortical bone thickness measurement: can it predict bone mineral density in the pediatric population?

Uğur Ufuk Işın[1]

Emin Çakmakçı[2]

Ayşe Derya Buluş[3]

Yüksel Yaşartekin[3]

Öznur Ünal[2]

Onur Dirican[4]

Abbas Ali Husseini[5]

[1]University of Health Sciences Türkiye, Atatürk Sanatorium Training and Research Hospital, Clinic of Pediatrics, Ankara, Türkiye

[2]University of Health Sciences Türkiye, Atatürk Sanatorium Training and Research Hospital, Clinic of Radiology, Ankara, Türkiye

[3]University of Health Sciences Türkiye, Atatürk Sanatorium Training and Research Hospital, Clinic of Pediatric Endocrinology, Ankara, Türkiye

[4]İstanbul Gelişim University, Vocational School of Health Services, Department of Pathology Laboratory Techniques, İstanbul, Türkiye

[5]İstanbul Gelişim University, Life Science and Biomedical Engineering Application and Research Center, İstanbul, Türkiye

PURPOSE

To explore sonographic cortical bone thickness (CoT) as a potential indicator of bone mineral density (BMD) measured by dual-energy X-ray absorptiometry for screening and diagnosing pediatric osteoporosis.

METHODS

A prospective study included 41 osteopenic or osteoporotic patients and 52 healthy children. Radius cortical thickness (R-CoT), tibial cortical thickness (T-CoT), and second metatarsal cortical thickness (M-CoT) were measured by B-mode ultrasound; CoT values were compared between groups and the correlation between BMD and CoT was examined.

RESULTS

There were no significant differences in R-CoT ($P = 0.433$), T-CoT ($P = 0.057$), and M-CoT ($P = 0.978$) values between the patient and control groups. No significant correlations were found between BMD T-scores and R-CoT ($r = -0.073$, $P = 0.490$), T-CoT ($r = -0.154$, $P = 0.141$), and M-CoT ($r = 0.047$, $P = 0.657$) values.

CONCLUSION

Sonographic CoT values in children do not correlate with BMD values. Unlike in adults, sonographic CoT measurements do not appear to have a role in assessing BMD in the pediatric population.

KEYWORDS

Osteoporosis, children, bone, density, cortical, thickness, ultrasonography

Corresponding author: Emin Çakmakçı

E-mail: em_sel74@hotmail.com

Osteoporosis, once primarily associated with the elderly, is increasingly being recognized as a concern in the pediatric population.[1] Various factors, such as genetics, diet, physical activity, medications, and the presence of chronic illnesses compromising bone strength, influence the bone health of children.[2,3] Untreated reduction in bone mass can result in deformities and negatively impact quality of life, potentially leading to long-term consequences.[4] Therefore, it is imperative to identify children with osteoporosis or those at high risk of developing it.

Pediatric osteoporosis is defined by the International Society of Clinical Densitometry using two criteria. The first criterion is a "low bone mineral content or bone mineral density (BMD)," characterized by a BMD Z-score of ≤-2. The second criterion is the "presence of a clinically significant fracture history," involving at least one long bone fracture in the lower extremity, at least two long bone fractures in the upper extremity, or a vertebral compression fracture.[5] Dual-energy X-ray absorptiometry (DXA) stands as the standard reference tool for assessing pediatric BMD.[6] However, alternative radiology tools, such as the bone health index,[7] direct radiography,[8] computed tomography (CT),[6] magnetic resonance imaging (MRI),[9] and ultrasonography[10] are also recommended for diagnosing and monitoring pediatric

osteoporosis. Nevertheless, each of these alternative methods presents drawbacks, and current recommendations emphasize the need for novel modalities in assessing osteoporosis in children.[1]

Children with osteoporosis may present with a history of recurrent fractures, deformities, or back pain.[11] An investigation for osteoporosis in children is warranted in the presence of significant risk factors, incidentally-detected fractures, or fractures that develop after minor trauma.[2,3] Given the increasing prevalence of osteoporosis in the pediatric age group, there is a growing demand for a diagnostic imaging modality that is easy to apply, safe, and cost-effective. In this context, ultrasonography has garnered attention from researchers due to its user-friendliness, affordability, and lack of ionizing radiation. As part of these efforts, ultrasonographic parameters, such as the speed of sound and broadband ultrasonography attenuation, have been explored.[12] A recent study in adults demonstrated a significant correlation between cortical bone thickness (CoT) measurements obtained through B-mode ultrasonography and BMD values determined via DXA.[13] However, the applicability and validity of this method in the pediatric population remain unknown.

Accordingly, the focus of our investigation lies in elucidating the relationship between CoT, as measured using B-mode ultrasonography, and BMD measurements obtained through DXA. Additionally, we aim to assess the utility and validity of sonographic CoT measurements as a screening and diagnostic tool for pediatric osteoporosis. Concurrently, within the scope of our study, it is imperative to highlight the significance of cortical thickness in relation to spatial resolution in ultrasonography. Specifically, we elucidate that when cortical thickness exceeds the inherent spatial resolution, a distinct demarcation can be observed in which both the external and

---

Main points

- In contrast to adults, the present study concluded that sonography does not play a role in assessing pediatric osteoporosis.

- There is no correlation between sonography-measured cortical thickness values and dual-energy X-ray absorptiometry-derived bone mineral density.

- Effective evaluation of pediatric patients requires considering a combination of objective radiological and/or biochemical data, clinical risk factors, and the limitations of pediatric osteoporosis criteria.

---

internal surfaces of the bone cortex manifest as separate, luminous lines within the ultrasound image. Conversely, when cortical thickness falls below the spatial resolution threshold, echoes from these surfaces overlap, resulting in a merged representation as a single conspicuously bright line within the ultrasound image. Notably, we emphasize that, in this study, echo thickness serves as a surrogate measure for cortical thickness.

Methods

Study design and recruitment of participants

This prospective cohort study, conducted from March to May 2023, included pediatric patients admitted to the endocrinology outpatient clinic with bone pain suggestive of osteoporosis. Ethics Committee approval was received for this study from the IRB Atatürk Sanatorium Training and Research Hospital (March 8, 2023, 2012-KAEK-15/2637). The parents or legal guardians of the participant children received a detailed explanation about the study, and written informed consent was obtained. The study was designed and conducted according to relevant ethical regulations and was performed following the Declaration of Helsinki and its later amendments.

Routine calcium level measurement was performed. Among the patients, 93 were identified as having hypocalcemia and underwent BMD and sonographic CoT measurements. Exclusion criteria included patients older than 18, refusal of BMD measurements and/or ultrasonography, missing data or laboratory results, and the presence of skeletal dysplasia or cerebral palsy. The patient group (n = 41) comprised children with osteopenia or osteoporosis based on BMD results, while the control group (n = 52) included healthy children.

Data collection

Participants' age and sex information, BMD measurement results, and CoT values were measured by ultrasonography. Blood samples were acquired from the antecubital vein to measure calcium, phosphorus, magnesium, alkaline phosphatase, parathyroid hormone, and 25-hydroxyvitamin D3 levels.

In this investigation, the Toshiba Aplio 500 device, along with its 14L5 frequency probe, was employed. During the examination, the ultrasound frequency utilized was 14 MHz, with a dynamic range set at 65, a frame rate of 6, and a single focus adjusted within the range of 45%–50%.

Bone mineral density measurement

Following blood sampling, BMD (mg/cm²) measurements were performed from the femoral neck and lumbar spine (L1–L4 posteroanterior) in all participants using a DXA device (Explorer QDR series; USA). The T-scores were calculated based on reference BMD datasets of Turkish children aged 6–18 years. T-score results were interpreted according to the World Health Organization criteria[14] as follows: normal (T-score ≥−1.0), osteopenia (−2.5< T-score <−1.0), and osteoporosis (T-score ≤−2.5).

Sonographic cortical bone thickness measurement

The ultrasonographic CoT values of each participant were measured using the same ultrasound device (Toshiba Aplio 500; Japan) and transducer. The measurements were taken by the same radiologist from the radius, tibia, and anterior cortical areas of the second metatarsal head of the non-dominant extremity. The radius cortical thickness (R-CoT) was measured 2 cm proximal to the radiocarpal joint, the tibial cortical thickness (T-CoT) was measured 2 cm proximal to the medial malleolus to the joint level, and the metatarsal cortical thickness (M-CoT) was measured 2 cm proximal to the second metatarsophalangeal joint. The B-mode ultrasonography images were adjusted for preset and gray scale settings. After achieving optimal focus and zoom settings perpendicular to the relevant bone cortex axis, the outer and inner starting and ending points of the first and most echogenic linear lines of the bone cortex were measured in millimeters (Figure 1).

Statistical analysis

The statistical analysis was performed using SPSS version 25.0 (IBM, USA). A p-value less than 0.05 was considered statistically significant for all analyses. The Shapiro–Wilk test was employed to assess whether continuous variables followed a normal distribution. Continuous variables were reported as mean ± standard deviation or median (first-to-third quartile) based on their distribution, while categorical variables were reported as relative frequency. Normally distributed continuous variables were analyzed with the Student's t-test, while non-normally distributed variables were analyzed using the Mann–Whitney U test. The chi-square test was used to analyze categorical variables. The Spearman correlation coefficient was calculated to assess the association between BMD and bone CoT.

Figure 1. The cortical bone thickness measurement.

Results

The mean age of the patient group was 11.25 ± 2.76 and 12.04 ± 3.03 in the control group. The female-to-male ratios were 28:13 and 30:22 in the patient and control groups, respectively. There was no significant difference between the two groups in terms of age and sex distribution ($P = 0.197$ and $P = 0.405$, respectively). There was no significant difference between the R-CoT ($P = 0.433$), T-CoT ($P = 0.057$), and M-CoT ($P = 0.978$) values of the patient and control groups. As expected, the median T-score ($P < 0.001$) and median calcium level ($P = 0.014$) of the patient group were significantly lower than the control group, while the alkaline phosphatase level was significantly higher ($P = 0.002$) (Table 1).

We also investigated the correlation between the T-score and CoT measured by sonography. No significant correlation was detected between T-score values and R-CoT ($r = -0.073$, $P = 0.490$), T-CoT ($r = -0.154$, $P = 0.141$), and M-CoT ($r = 0.047$, $P = 0.657$) (Table 2).

Discussion

The present study's results underscore the limited utility of sonography in assessing pediatric osteoporosis, despite its use indicating intriguing relationships in adult studies. This discrepancy may be attributed to fundamental disparities between pediatric and adult osteoporosis.[2] It is paramount to acknowledge that nearly 95% of skeletal size, bone mass, and muscle mass are acquired before the age of 18.[15] While rare, untreated osteoporosis during childhood can carry adverse repercussions into adulthood.

Pediatric osteoporosis often arises as a complication of underlying diseases or as a side effect of medications.[16] Furthermore, given the higher frequency of fractures resulting from behavioral factors in childhood, distinguishing whether these fractures are pathological or a natural consequence of high-energy trauma becomes more challenging.[17] Nevertheless, specific scenarios warrant suspicion of pediatric osteoporosis. For instance, any child presenting with back pain should undergo evaluation for occult vertebral fractures through lateral spine X-rays to exclude osteoporosis. Similarly, pediatric osteoporosis should be considered when fractures occur following low-energy trauma.[2,3] In addition to these challenges, the limitations of pediatric osteoporosis criteria necessitate the comprehensive evaluation of pediatric patients, incorporating objective radiological and/or biochemical data, as well as clinical risk factors. Identifying patients at high risk for low BMD or with low BMD is critical to administering effective treatment. However, presently, DXA remains the predominant modality for assessing pediatric osteoporosis, highlighting the need for radiological tools that can overcome its limitations, such as cost, availability, and radiation exposure. This study investigated whether a correlation could be established between sonography-measured CoT values and BMD results obtained via DXA. However, our findings did not establish a significant relationship between these two parameters.

Ultrasonography presents potential as an alternative to BMD assessment, with the most explored ultrasonographic parameters being speed-of-sound (or ultrasound velocity) and broadband ultrasonography attenu-

Table 1. Summary of patient characteristics and measurements with regard to groups

		Groups		
	Total	Patients (n = 41)	Controls (n = 52)	P
Age	11.69 ± 2.93	11.25 ± 2.76	12.04 ± 3.03	0.197
Sex				
Female	58 (62.37%)	28 (68.29%)	30 (57.69%)	0.405
Male	35 (37.63%)	13 (31.71%)	22 (42.31%)	
Bone mineral density, T-score	−0.76 (−1.95–0.42)	−2.12 (−2.81 to −1.36)	0.38 (−0.48–1.05)	**<0.001**
Normal (T-score ≥ −1.0)	52 (55.91%)	0 (0.00%)	52 (100.00%)	
Osteopenia (−2.5 < T-score <−1.0)	29 (31.18%)	29 (70.73%)	0 (0.00%)	**<0.001**
Osteoporosis (T-score ≤−2.5)	12 (12.90%)	12 (29.27%)	0 (0.00%)	
Bone cortical thickness				
Radius	0.6 (0.5–0.7)	0.6 (0.5–0.8)	0.6 (0.5–0.7)	0.433
Tibia	0.6 (0.6–0.8)	0.7 (0.6–0.8)	0.6 (0.6–0.7)	0.057
Second metatarsal	0.5 (0.4–0.6)	0.5 (0.4–0.6)	0.5 (0.45–0.6)	0.978
Calcium	9.7 (9.3–10.1)	9.6 (8.9–10.0)	9.9 (9.5–10.1)	**0.014**
Phosphorus	4.6 (4.1–4.9)	4.8 (4.2–5.1)	4.6 (4.05–4.8)	0.089
Magnesium	2.0 (1.9–2.1)	2.0 (1.9–2.1)	2.0 (1.9–2.1)	0.850
Alkaline phosphatase	207 (124–249)	227 (167.5–430)	183 (115.5–224.5)	**0.002**
Parathyroid hormone	78.0 (66.2–95.05)	82.2 (70.5–115.6)	76.5 (61.0–86.0)	0.104
25-hydroxyvitamin D3	14.0 (9.2–18.55)	15.0 (8.1–22.0)	13.8 (10.2–16.6)	0.683

Data are given as mean ± standard deviation or median (first-to-third quartile) for continuous variables according to normality of distribution and as a frequency (percentage) for categorical variables.

Table 2. Correlations between bone mineral density and cortical thickness measurements

		Bone cortical thickness		
		Radius	Tibia	Second metatarsal
Bone mineral density, T-score	r	−0.073	−0.154	0.047
	p	0.490	0.141	0.657

r, Spearman's correlation coefficient.

ation.[12] Previous studies have demonstrated significant correlations between CoT determined by CT or direct X-ray and bone mass and BMD.[18,19] In line with these investigations, Gokcek et al.[13] noted that ultrasonographic R-CoT and T-CoT values played prognostic roles in predicting patients with abnormal T-scores measured by DXA. Nevertheless, our study found no significant differences in R-CoT, T-CoT, and M-CoT values between the patient and control groups. While there is insufficient evidence in the literature to make recommendations regarding the utility of ultrasonography in predicting fractures in children,[16] Hartman et al.[12] suggest a significant positive correlation between lumbar DXA and radius speed-of-sound, asserting that ultrasound evaluation of the radius and tibia can yield results comparable to DXA for screening pediatric osteoporosis.

The management of pediatric osteoporosis involves two crucial stages: identifying high-risk children for low BMD and diagnosing osteoporosis.[2,3,6] In children with suspected osteoporosis who present with fractures, deformities, or bone pain, detailed laboratory examinations are conducted for diagnostic purposes. Radiological evaluations, including DXA and, if necessary, conventional lateral spine radiographs, are ordered for these patients. Despite its various limitations, DXA is still considered the reference standard technique for assessing bone quality and detecting pediatric osteoporosis due to its standardized results. However, the use of DXA in the pediatric population is inherently limited due to ionizing radiation exposure.[16] Additionally, DXA measurements may underestimate BMD in children with short stature or delayed puberty. Furthermore, the effects of growth retardation are not considered for children younger than 5 years.[20] Various factors, including movement during scans, scoliosis, body size, ethnicity, bone age, and pubertal development, can influence BMD results.[1] Moreover, DXA lacks the ability to distinguish between trabecular and cortical bone or provide information about bone geometry.[1] Despite its status as the primary modality for osteoporosis investigation, DXA may not be the most suitable tool for detecting changes in bone mass.[1]

The limitations of DXA have driven researchers to explore alternative or supplementary diagnostic tools. The bone health index measured via radiogrammetry,[7] direct radiography for detecting occult vertebral fractures,[8] quantitative CT (qCT),[6] MRI,[9] and quantitative ultrasonography (qUS)[10] are among the radiological tools used or rec-

ommended for pediatric osteoporosis diagnosis. However, the bone health index has a high false positive rate and a weak correlation with BMD,[3] while direct X-ray serves as a semi-quantitative method used solely to identify occult vertebral fractures and confirm suspected osteoporosis.[8] Alternatively, MRI and qCT offer the advantage of separately evaluating cortical and trabecular bone, with qCT having proved to be an alternative diagnostic tool in children with severe scoliosis or joint contractures.[1,3] In comparison to DXA, qCT BMD provides more valuable information about bone features, enhancing our understanding of skeletal defects associated with fracture risk.[1] Nonetheless, the higher radiation exposure associated with qCT limits its use in pediatric osteoporosis diagnosis.[1] Despite multiple studies exploring these modalities and their listed advantages, DXA remains the gold standard in the majority of healthcare centers. Ultrasonography stands out due to its lack of ionizing radiation emission, portability, non-invasiveness, affordability, ability to establish normative values for pediatric patients without radiation concerns, and capacity to measure other bone properties such as elasticity, microarchitecture, and thickness.[12,21,22] However, while qUS is generally accepted for osteoporosis screening in adults, given its effectiveness as a fracture risk indicator in postmenopausal women,[16,23] the same cannot be stated for its use in the pediatric population.[10,16]

The potential for pediatric osteoporosis and an increased fracture risk can be influenced by factors such as genetic disorders, lifestyle, chronic diseases, specific medications, calcium and vitamin D intake, and weight-bearing exercise.[3] While many pediatricians may not opt for further osteoporosis examination in children who experience bone fractures due to severe trauma, fractures resulting from mild trauma merit investigation for osteoporosis. However, reluctance to undergo DXA measurements, primarily due to radiation exposure, necessitates further studies to evaluate different modalities for their role in detecting osteoporosis. Achieving this requires screening tests that are minimally affected by external factors, exhibit high sensitivity and specificity, are easy to administer, have minimal or no side effects, and, most importantly, possess a high level of applicability and validity in the pediatric population. Despite the demonstrated utility of sonographic CoT measurement in adults,[13] there is a notable absence of evidence supporting its effectiveness in pediatric patients, as confirmed by this study.

Limitations of the study encompass various aspects that warrant consideration. First, a noteworthy limitation is the preponderance of osteoporotic patients within the study's patient group, constituting 12 of 41 patients. This preeminence of osteoporosis in the patient cohort has implications for the study's generalizability and the interpretation of its findings. Notably, osteoporosis signifies an advanced stage of bone loss compared to osteopenia, reflecting the severity of bone health issues within this subset of pediatric patients. This skewed patient group composition introduces a potential bias when drawing comparisons between the patient and the healthy control groups. Given the majority of osteoporotic patients in the study, any findings pertaining to CoT and its association with BMD may be disproportionately influenced by the characteristics specific to osteoporotic patients, potentially deviating from the broader pediatric population's characteristics according to varying degrees of bone health.

The above imbalance raises concerns about the applicability and representativeness of the study's conclusions for the wider pediatric population. Moreover, it may have ramifications for the statistical analyses and correlations involving CoT and BMD, as the patient group's composition may skew the results towards osteoporotic traits. Consequently, it is imperative to acknowledge this patient group's predominance as a notable limitation when discussing the study's findings and their relevance to the pediatric population as a whole.

Furthermore, it is crucial to acknowledge that all patients included in the study exhibited symptoms and laboratory findings that prompted suspicion of osteoporosis, leading to the recommendation of DXA measurements. Consequently, there exists the possibility that some bone properties of the control group may have exhibited similarities to those of the patients, potentially introducing confounding factors into the comparisons. Additionally, despite efforts to maintain similarities in age and sex distribution between the groups, factors such as body size, weight, physical activity, and other uncontrolled variables may have significantly influenced bone thickness measurements. Future studies could benefit from the inclusion of a control group comprised exclusively of volunteers, which may help mitigate potential biases.

Moreover, the grouping of osteopenic and osteoporotic patients for comparative analyses was necessitated by the limited number

of patients in each category. Although separate statistical analyses were conducted for subgroups, these efforts did not yield significant differences or noteworthy findings. This limitation underscores the challenge of conducting more nuanced analyses due to sample size constraints.

Finally, the study's primary focus on investigating the relationship between ultrasonographic CoT and BMD led to the inclusion of a restricted number of variables. Detailed data pertaining to patients' clinical characteristics, the specific type and etiology of osteoporosis, the presence of chronic rheumatological diseases, fracture history, occult bone fractures, and long-term follow-up complications were beyond the study's scope. These omissions highlight the need for future research endeavors to consider a broader array of factors and variables in the investigation of pediatric osteoporosis.

In conclusion, it is evident from our results that sonographic R-CoT, T-CoT, and M-CoT measurements did not differ in child patients with osteopenia or osteoporosis compared to healthy controls in our population. Similarly, no correlations were found between sonographic R-CoT, T-CoT, M-CoT values, and BMD T-scores obtained via DXA. In contrast to the adult population, sonographic CoT measurements appeared to be unassociated with BMD in children, and, accordingly, these measures cannot be used to assess BMD in the pediatric population.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Marrani E, Giani T, Simonini G, Cimaz R. Pediatric osteoporosis: diagnosis and treatment considerations. *Drugs*. 2017;77(6):679-695. [CrossRef]

2. Steffey CL. Pediatric osteoporosis. *Pediatr Rev*. 2019;40(5):259-261. [CrossRef]

3. Ciancia S, van Rijn RR, Högler W, et al. Osteoporosis in children and adolescents: when to suspect and how to diagnose it. *Eur J Pediatr*. 2022;181(7):2549-2561. [CrossRef]

4. Gordon CM, Zemel BS, Wren TAL, et al. The determinants of peak bone mass. *J Pediatr*. 2017;180:261-269. [CrossRef]

5. Gordon CM, Leonard MB, Zemel BS; International Society for Clinical Densitometry. 2013 Pediatric Position Development Conference: executive summary and reflections. *J Clin Densitom*. 2014;17(2):219-224. [CrossRef]

6. Medeleanu M, Vali R, Sadeghpour S, Moineddin R, Doria AS. A systematic review and meta-analysis of pediatric normative peripheral quantitative computed tomography data. *Bone Rep*. 2021;15:101103. [CrossRef]

7. Rosholm A, Hyldstrup L, Bæksgaard L, Grunkin M, Thodberg HH. Estimation of bone mineral density by digital X-ray radiogrammetry: theoretical background and clinical testing. *Osteoporos Int*. 2001;12(11):961-969. [CrossRef]

8. Alqahtani FF, Offiah AC. Diagnosis of osteoporotic vertebral fractures in children. *Pediatr Radiol*. 2019;49(3):283-296. [CrossRef]

9. Di Iorgi N, Maruca K, Patti G, Mora S. Update on bone density measurements and their interpretation in children and adolescents. *Best Pract Res Clin Endocrinol Metab*. 2018;32(4):477-498. [CrossRef]

10. Baroncelli GI. Quantitative ultrasound methods to assess bone mineral status in children: technical characteristics, performance, and clinical application. *Pediatr Res*. 2008;63(3):220-228. [CrossRef]

11. Saraff V, Högler W. Endocrinology and adolescence: osteoporosis in children: diagnosis and management. *Eur J Endocrinol*. 2015;173(6):185-197. [CrossRef]

12. Hartman C, Shamir R, Eshach-Adiv O, Iosilevsky G, Brik R. Assessment of osteoporosis by quantitative ultrasound versus dual energy X-ray absorptiometry in children with chronic rheumatic diseases. *J Rheumatol*. 2004;31(5):981-985. [CrossRef]

13. Gokcek A, Karabay EO, Yegin MA, Cakmakci E, Gencer S. Can ultrasonographic measurement of bone cortical thickness predict osteoporosis? *Acad Radiol*. 2023;30(3):516-527. [CrossRef]

14. Camacho PM, Petak SM, Binkley N, et al. American Association of Clinical Endocrinologists/American College of Endocrinology Clinical Practice Guidelines for the Diagnosis and Treatment of Postmenopausal Osteoporosis - 2020 Update Executive Summary. *Endocr Pract*. 2020;26(5):564-570. [CrossRef]

15. Baxter-Jones AD, Faulkner RA, Forwood MR, Mirwald RL, Bailey DA. Bone mineral accrual from 8 to 30 years of age: an estimation of peak bone mass. *J Bone Miner Res*. 2011;26(8):1729-1739. [CrossRef]

16. Wang KC, Wang KC, Amirabadi A, et al. Evidence-based outcomes on diagnostic accuracy of quantitative ultrasound for assessment of pediatric osteoporosis - a systematic review. *Pediatr Radiol*. 2014;44(12):1573-1587. [CrossRef]

17. Larsen AV, Mundbjerg E, Lauritsen JM, Faergemann C. Development of the annual incidence rate of fracture in children 1980-2018: a population-based study of 32,375 fractures. *Acta Orthop*. 2020;91(5):593-597. [CrossRef]

18. Baroncelli GI, Federico G, Bertelloni S, et al. Assessment of bone quality by quantitative ultrasound of proximal phalanges of the hand and fracture rate in children and adolescents with bone and mineral disorders. *Pediatr Res*. 2003;54(1):125-136. [CrossRef]

19. Sadat-Ali M, Elshaboury E, Al-Omran AS, Azam MQ, Syed A, Gullenpet AH. Tibial cortical thickness: a dependable tool for assessing osteoporosis in the absence of dual energy X-ray absorptiopmetry. *Int J Appl Basic Med Res*. 2015;5(1):21-24. [CrossRef]

20. Bachrach LK, Gordon CM; Section on Endocrinology. Bone densitometry in children and adolescents. *Pediatrics*. 2016;138(4):e20162398. [CrossRef]

21. Hartman C, Hino B, Lerner A, et al. Bone quantitative ultrasound and bone mineral density in children with celiac disease. *J Pediatr Gastroenterol Nutr*. 2004;39(5):504-510. [CrossRef]

22. Halaba ZP, Bursa J, Kaplon UK, Pluskiewicz W, Marciniak S, Drzewiecka U. Phalangeal quantitative ultrasound measurements in former pre-term children aged 9-11 years. *Br J Radiol*. 2007;80(954):401-405. [CrossRef]

23. Christoforidis A, Economou M, Papadopoulou E, et al. Comparative study of dual energy X-ray absorptiometry and quantitative ultrasonography with the use of biochemical markers of bone turnover in boys with haemophilia. *Haemophilia*. 2011;17(1):217-222. [CrossRef]

Establishment of local diagnostic reference levels for computed tomography with cloud-based automated dose-tracking software in Türkiye

🆔 Gökhan Kahraman
🆔 Kemal Murat Haberal
🆔 Ahmet Muhteşem Ağıldere

Başkent University Faculty of Medicine, Department of Radiology, Ankara, Türkiye

PURPOSE

The purpose of this study is to establish local diagnostic reference levels (LDRLs) for computed tomography (CT) procedures using cloud-based automated dose-tracking software.

METHODS

The study includes the dose data obtained from a total of 104,272 examinations performed on adult patients (>18 years) using 8 CT scanners over 12 months. The protocols included in our study were as follows: head CT without contrast, cervical spine CT without contrast, neck CT with contrast, chest CT without contrast, abdomen–pelvis CT without contrast, lumbar spine CT without contrast, high-resolution computed tomography (HRCT) of the chest, and coronary CT angiography (CTA). Dose data were collected using cloud-based automatic dose-tracking software. The 75th percentiles of the distributions of the median volume CT dose index (CTDIvol) and dose length product (DLP) values were used to determine the LDRLs for each protocol. The LDRLs were compared with national DRLs (NDRLs) and DRLs set in other countries. Inter-CT scanner variability, which is a measure of how well clinical practices are standardized, was determined for each protocol. Median values for each protocol were compared with the LDRLs for dose optimization in each CT scanner.

RESULTS

The LDRLs (for DLP and CTDIvol, respectively) were 839 mGy.cm and 41.2 mGy for head CT without contrast, 530.6 mGy.cm and 19.8 mGy for cervical spine CT without contrast, 431.9 mGy.cm and 15.5 mGy for neck CT with contrast, 364.8 mGy.cm and 9.3 mGy for chest CT without contrast, 588.9 mGy.cm and 11.2 mGy for abdomen–pelvis CT without contrast, 713 mGy.cm and 24.3 mGy for lumbar spine CT without contrast, 326 mGy.cm and 9.5 mGy for HRCT, and 642.3 mGy.cm and 33.4 mGy for coronary CTA. The LDRLs were comparable to or lower than NDRLs and DRLs set in other countries for most protocols. The comparisons revealed the need for immediate initiation of an optimization process for CT protocols with higher dose distributions. Furthermore, protocols with high inter-CT scanner variability revealed the need for standardization.

CONCLUSION

There is a need to update the NDRLs for CT protocols in Türkiye. Until new NDRLs are established, local institutions in Türkiye can initiate the optimization process by comparing their dose distributions to the LDRLs established in our study. Automated dose-tracking software can play an important role in establishing DRLs by facilitating the collection and analysis of large datasets.

KEYWORDS

Computed tomography, diagnostic reference level, dose-tracking software, ionizing radiation, radiation dose

Corresponding author: Gökhan Kahraman

E-mail: gokhankahraman1@outlook.com

Computed tomography (CT) is frequently preferred by clinicians, as it provides rapid and non-invasive imaging of patients and makes significant contributions to patient management. CT scanners have become widely used since their introduction, and the frequency of CT scans has expanded significantly, particularly in the last few decades. Thus, CT has become the imaging method with the highest proportion of radiation exposure among imaging methods. Despite accounting for only 10% of radiological procedures, it accounts for approximately 62% of the collective effective dose (ED).[1]

The International Commission on Radiological Protection (ICRP), which first introduced the term "diagnostic reference level (DRL)" in 1996, recommended the establishment of DRLs in Publication 103 to cope with increased medical exposure and optimize radiation dose.[2] Publication 135 provides clarification on the term DRL and how DRLs should be established.[3] DRLs provide periodic monitoring of radiation dose levels. The DRL process has made significant contributions to radiation dose optimization in many countries.[4]

The DRL is established as the 75th percentile (third quartile) of median dose values for each CT protocol. National DRLs (NDRLs) represent the entire country, whereas local DRLs (LDRLs) represent a group of healthcare facilities in an area. In theory, LDRL should not exceed NDRL, and if it does, the dose optimization process should begin immediately. The DRL process is completed in a short time with automated dose-tracking software that facilitates the collection and analysis of dose data. These software programs allow for the DRLs to be updated more frequently, contributing significantly to the process of dose optimization. Using these software packages, health center dose data can be easily monitored, and the optimization process can be started immediately when dose values exceed reference DRLs.[5]

The first DRL study in Türkiye was conducted by Ataç et al.[6] NDRLs are yet to be established for many CT protocols, and existing NDRLs need to be updated as part of the ICRP recommendations. LDRLs may be set for procedures for which no NDRLs are available, according to the ICRP.[3]

In this context, in our study, LDRLs were determined for 8 CT protocols using cloud-based automated dose tracking software to initiate the dose optimization process in our institution's CT scanners and to contribute to the national CT dose optimization efforts in Türkiye.

Methods

Computed tomography scanners and protocols

In total, the data of 104,272 doses were collected from CT examinations of adult patients over 18 years of age, performed between January 1, 2020, and December 31, 2020, using 8 CT scanners in 5 university hospitals. Different models of CT scanners from the three major CT manufacturers (Siemens, Toshiba, and GE) were used in the study. Details about the CT scanners are presented in Table 1. The CT protocols were as follows: head CT without contrast, cervical spine CT without contrast, neck CT with contrast, chest CT without contrast, abdomen–pelvis CT without contrast, lumbar spine CT without contrast, high-resolution CT (HRCT) of the chest, and coronary CT angiography (CTA). For the coronary CTA protocol, the data were collected from a total of 4 CT scanners. For the other 7 protocols, data were collected from all CT scanners, and single-phase acquisitions were included in the study. Automatic exposure control was used in all protocols. Quality control tests of all the CT scanners were completed in December 2019. For all protocols, the post-exposure volume CT dose index (CTDIvol) values provided by CT scanners were confirmed by direct measurements performed on standard polymethylmethacrylate CT phantoms with a diameter of 16 and 32 cm. A 16-cm phantom was used for head CT and a 32-cm phantom for other protocols. This study was approved by the institutional review board and Başkent University Institutional Review Board and Ethics Committee (project no: KA18/206, date of approval, 26/06/2018) and the need to obtain informed consent was waived.

Data collection

Before collecting the dose data, a standard protocol nomenclature was determined to ensure correct analysis of the study data. The dose data were collected with Teamplay (Siemens Healthineers, Erlangen, Germany), a cloud-based automated dose-tracking software. This software collects data from Radiation Dose Structured Report on Digital Imaging and Communications in Medicine which is the international standard primarily used for storing and transmitting medical images. The data collected were as follows; anonymous ID for each patient, patient age and sex, examination date, protocol name, anatomical site, CTDIvol, dose length product (DLP), health center name, CT scanner name, scanning parameters, and CTDI phantom. Anonymized data were transferred to Excel (Microsoft, Redmond, Wash.) file. Dose data from duplicated exams, exams including additional acquisitions, exams performed under a name other than the standardized protocol name, scout acquisitions, and bolus tracking acquisitions were not included. The correct use of phantoms was checked for

Table 1. Details of CT scanners

No	Manufacturer	Model	Number of detector rows	Year of installation	Iterative reconstruction
1	Siemens	Somatom Force	2 x 192	2018	Yes
2	Siemens	Somatom go.All	64	2018	Yes
3	Siemens	Sensation 64	64	2007	No
4	Siemens	Somatom Definition AS 64	64	2017	Yes
5	Toshiba	Aquilion CX	64	2012	No
6	Siemens	Sensation 16	16	2012	No
7	GE	BrightSpeed Elite Select 16	16	2012	No
8	GE	CT580 RT	16	2014	No

CT, computed tomography.

each examination. ICRP recommends weight standardization (mean 70 ± 5 kg) for adult patients if the number of patients is less than 50.[3] Since weight standardization could not be performed, dose data obtained from protocols with more than 50 patients were included in the study.

Statistical analysis

The 25th, 50th, and 75th percentiles of the distribution of medians for each protocol were calculated using SPSS v.27.0 (IBM Corp., Armonk, United States). The LDRLs for CTDIvol and the DLP for each protocol were determined as the 75th percentile (third quartile) of the distribution of medians according to ICRP recommendations.[3] Descriptive comparisons were made with NDRLs and DRLs set in other countries. Statistical comparisons were not performed because of the methodological variations between countries and insufficient data. The interquartile range (IQR) for each protocol was divided into the median (50th percentile) to determine inter-CT scanner variability. The median values for each protocol were compared with LDRLs for dose optimization in each CT scanner.

Results

Of the 104,272 CT exams performed, 51.6% were on male patients, while 48.4 % were on female patients. The mean age of the patients was 58.4 years, with a range of 18 to 103. Chest CT without contrast was the most common protocol (n = 50984, 48.9%), whereas cervical spine CT without contrast was the least common (n = 1270, 1.2%) (Table 2). Since weight standardization could not be performed, the dose data from a total of 167 CT examinations from protocols with fewer than 50 examinations were excluded from the study.

Table 3 shows the LDRLs, 25th and 50th percentiles, and IQR/50th percentile values. Table 4 compares the LDRLs, NDRLs, and DRLs set in other countries. Among the compared DRLs, NDRLs from the UK, EU, and Switzerland were established based on clinical indications, while DRLs from other countries were set based on anatomical location.[6-14] Four of the eight protocols–head CT, chest CT, HRCT, and abdomen CT–were comparable to NDRLs. The LDRLs were comparable to or lower than NDRLs for most comparable protocols. For all four comparable protocols, the LDRLs for CTDIvol were lower than the NDRLs. The LDRLs for DLP were higher than the comparable two protocols (abdomen and chest CT) from the NDRLs.[6] The LDRLs for

CTDIvol and DLP were lower than the NDRLs of the US, Japan, and Canada for all comparable protocols.[7-9] The LDRLs were comparable or lower than the UK, EU, German, Swiss, and Korean NDRLs for most comparable protocols.[10-14]

In Figure 1, the median values of the doses obtained from all CT scanners for each protocol are compared with the LDRLs. Median dose values were higher than LDRLs in CT 4, CT 5, and CT 7 scanners for head CT, CT 5 and CT 7 scanners for cervical spine CT, CT 5 and CT 6 scanners for neck CT, CT 5 and CT 8 scanners for chest CT, CT5 scanner for abdomen–pelvis CT, CT 2 and CT 5 scanners for lumbar spine CT, CT 5, CT 7 and CT 8 scanners for HRCT, and CT 5 scanner for coronary CTA.

Discussion

The ICRP introduced the term DRL in 1996.[15] The establishment of DRLs was the first step in the radiation dose optimization process. DRL enables units and hospitals to compare radiation doses to identify variations among them. Accordingly, it helps to maintain radiation doses at an acceptable

level and aids in their optimization. DRLs are not strict dose limits or concepts generated to establish legal standards but should be used to determine whether doses are high. The DRL process begins with the collection of dose data. The collected data is then plotted in a histogram and the 75th percentile of the histogram is determined as the DRL.[3] The DRL process has become popular in many countries, and the use of DRLs has led to a decrease in both radiation dose and the range of radiation doses, resulting in successful outcomes in radiation dose optimization.[4,16]

NDRLs represent the entire country, while LDRLs represent several healthcare facilities in an area. The first NDRLs in Türkiye were published in 2015 by Ataç et al.[6] NDRLs were established for single-phase CT protocols. The ICRP recommends that NDRLs be updated every three to five years. It is also recommended that the process of updating DRLs be both flexible and dynamic. Flexibility is necessary for procedures with limited data or where data can be obtained from only one or a few centers. Initial DRLs can be established using the limited data available before conducting more comprehensive DRL studies.

Table 2. Number of examinations

	CT 1	CT 2	CT 3	CT 4	CT 5	CT 6	CT 7	CT 8	Total	Percentage
Head	2662	1981	1030	748	2795	9353	951	86	19606	18.80%
Cervical spine	325	221	57	31	50	544	34	8	1270	1.22%
Neck	89	22	50	60	50	314	612	409	1606	1.54%
Chest	8703	8639	3000	4105	8395	9567	4456	4119	50984	48.90%
Abdomen	6085	3207	1030	542	1920	4596	3544	3645	24569	23.56%
Lumbar spine	303	165	50	174	100	594	52	17	1455	1.40%
HRCT	242	197	30	25	75	1818	166	67	2620	2.51%
Coronary CTA	946	-	430	686	100	-	-	-	2162	2.07%

CT, computed tomography; HRCT, high-resolution computed tomography; CTA, computed tomography angiography.

Table 3. Local diagnostic reference levels (75th percentile), 50th percentile, 25th percentile, and IQR/median values for CTDIvol (mGy) and DLP (mGy.cm)

	25th percentile		50th percentile		75th percentile (DRL)		IQR/median	
	CTDIvol	DLP	CTDIvol	DLP	CTDIvol	DLP	CTDIvol	DLP
Head	33.7	673.3	38.1	752.2	41.2	839	0.20	0.22
Cervical spine	10.7	250.8	16.2	379.4	19.8	530.6	0.56	0.74
Neck	7.9	236.3	9.3	287.7	15.5	431.9	0.82	0.68
Chest	4.3	160.3	6	230.6	9.3	364.8	0.83	0.89
Abdomen	6.5	319.5	8.1	426.7	11.2	588.9	0.58	0.63
Lumbar spine	14.7	384.7	19.6	536.6	24.3	713	0.49	0.61
HRCT	6.1	204.2	7.5	256.4	9.5	326	0.45	0.48
Coronary CTA	11.9	190.1	26.1	405.8	33.4	642.3	0.82	1.11

CTDIvol, volume computed tomography dose index; mGy, milligray; DLP, dose length product; mGy.cm, milligray centimeter; DRL, diagnostic reference level; IQR, interquartile range; HRCT, high-resolution computed tomography; CTA, computed tomography angiography.

Table 4. Comparison of the LDRLs with the NDRLs and DRLs set in other countries for CTDIvol (mGy) and DLP (mGy.cm)

		Head	Cervical spine	Neck	Abdomen	Chest	HRCT	Lumbar spine	Coronary CTA
LDRL (this study)	CTDIvol	41.2	19.8	15.5	11.2	9.3	9.5	24.3	33.4
	DLP	839	530.6	431.9	588.9	364.8	326	713	642.3
Türkiye (NDRL)[6]	CTDIvol	66.4	-	-	13.3	11.6	11.3	-	-
	DLP	810	-	-	204	289	283	-	-
US[7]	CTDIvol	57	28	20	20	15	-	-	-
	DLP	1011	602	572	1004	545	-	-	-
UK[10]	CTDIvol	47	16	-	10	8.5	8	-	-
	DLP	790	400	-	530	290	300	-	-
EU[11]	CTDIvol	48	17	-	9	9	-	-	25
	DLP	1386	495	-	874	364	-	-	459
Germany[12]	CTDIvol	60	20	15	15	10	3	10	-
	DLP	850	300	330	700	350	100	180	-
Japan[8]	CTDIvol	77	-	-	18	13	-	-	66
	DLP	1350	-	-	880	510	-	-	1300
Switzerland[13]	CTDIvol	51	17	16	11	7	-	-	-
	DLP	890	360	410	540	250	-	-	-
Korea[14]	CTDIvol	52.2	20.9	13.4	10.3	7.6	-	20.6	19.2
	DLP	969.8	508.7	597.1	558.5	324.2	-	738.5	326.9
Canada[9]	CTDIvol	82	-	-	18	14	-	-	-
	DLP	1302	-	-	874	521	-	-	-

LDRLs, local diagnostic reference levels; NDRLs, national diagnostic reference levels; DRLs, diagnostic reference levels; CTDIvol, volume computed tomography dose index; mGy, milligray; DLP, dose length product; mGy.cm, milligray centimeter; HRCT, high-resolution computed tomography; CTA, computed tomography angiography; US, United States; UK, United Kingdom; EU, European Union.

Figure 1. Boxplots of DLP distributions for CT protocols, performed on 8 CT scanners. The upper, central, and lower lines of each box correspond to the 25th, 50th, and 75th percentiles, respectively. The vertical red lines show the 75th percentiles of the distribution of the median values (LDRLs). DLP, dose length product; CT, computed tomography; LDRLs, local diagnostic reference levels; HRCT, high-resolution computed tomography; CTA, computed tomography angiography.

Furthermore, it has been noted that LDRLs can be set for procedures for which no NDRLs are available.[3] There is a need to update NDRLs and establish NDRLs for many other CT protocols. In this context, LDRLs were established in our study using automatic dose monitoring software to begin the dose optimization process in our institution's CT scanners by comparing them to the NDRLs and to contribute to the national dose optimization efforts. Until new NDRLs are established, local institutions in Türkiye can initiate the optimization process by comparing their dose distributions to the LDRLs established in our study.

The LDRLs were comparable to or lower than NDRLs and DRLs set in other countries for most comparable protocols. This situation can be attributed to the use of CT scanners or software with newer technology in our study. New CT technologies, including iterative reconstruction algorithms, automatic exposure control devices, new noise reduction techniques, and detectors with high quantum detective efficiency, can significantly reduce radiation dose.[17-19] The ICRP recommends updating DRLs with the use of new technologies and software.[3] The high LDRLs indicated that the optimization process

should be initiated immediately. The LDRLs for DLP were higher than the comparable two protocols (abdomen and chest CT) from NDRLs. Since DLP is related to scan length, this situation was attributed to the high scan length of the CT examinations performed in our institution. This problem can be solved by reducing the scan lengths.

IQR (Q3–Q1) is a measure of the distribution of data. Inter-CT scanner variability (in terms of IQR/median) is an indicator of the standardization of clinical practice for a particular protocol.[13] The lack of protocol standardization leads to wide variations in radiation, even within the same healthcare facility.[20] The high inter-CT scanner variability observed in chest and neck CT, as well as coronary CTA protocols, indicates the need for standardizing these protocols.

The median values (50th percentile) of doses obtained from CT scanners are considered as "achievable or typical doses".[3,21] If the median value of doses obtained from the CT scanner for a specific protocol exceeds the reference DRL, it indicates the need for dose optimization.[3] As an example from our study, CT scanners, including CT 4, CT 5, and CT 7, require optimization for head CT (Figure 1).

The findings of our study showed that there were significant dose variations between the CT scanners (Figure 1). Dose variations may result from scanners, scanning protocols, and radiographer training and experience.[22] Dose optimization can be improved by staff training.[23,24] Staff training should be provided to ensure proper collimation and the correct use of equipment, and appropriate scanning parameters should be established and continuously monitored.

Ataç et al.[6] reported that the response rate to questionnaires was lower than expected during the establishment of DRLs. It has been suggested that the use of systems that enhance inter-institutional communication, such as internet-based questionnaires, could be beneficial in dose studies.[6] The cloud-based nature of the software we used in our study facilitated access to dose data from different centers. The use of cloud-based dose-tracking systems in DRL studies can enable the easy acquisition of data from numerous CT scanners.

We observed significant variations in protocol names before collecting dose data. It should be ensured that each CT examination is performed with common names before establishing DRLs. The common nomenclature for each protocol is crucial for data

validation. Kanal et al.[7] highlighted that improper labeling of CT protocols could lead to problems with dose data; they suggested standardizing the protocol names according to Radlex terminology as a solution.[25] In our study, before collecting dose data, we standardized the CT protocol names through collaboration among radiologists, radiographers, and technologists. If the common protocol names were not used, the number of dose data would be reduced while the LDRLs would be increased due to the inclusion of dose data from multiphase examinations.

There are several limitations to our study, which was conducted using dose data obtained from three among four major CT scanner manufacturers (Philips was not available). Most of the dose data were obtained from head and chest CT, as well as abdomen-pelvis CT protocols. For other protocols, a smaller amount of dose data was available (Table 2). This study can be strengthened by including dose data from all CT scanner manufacturers over a longer period. LDRLs were established for single-phase protocols to enable comparison with NDRLs; however, DRLs for multiphase protocols, which constitute a significant portion of routine practice, were not established. Although we established DRLs for four protocols in addition to NDRL protocols, it is necessary to include many others, including multi-phase protocols, in future DRL studies.

DRLs are dose levels in radio-diagnostic practices for standard-sized patient groups. To ensure meaningful comparisons of DRLs, it is recommended that dose data from standard-sized patient groups be included when establishing DRLs.[15] However, in ICRP Publication 135, it is stated that if an automated data collection system is used, DRLs can be established using all dose data, and it may be possible to relax weight restrictions.[3] In our study, dose data were collected using the automated dose tracking software, Teamplay, including all dose data without weight standardization. Automated dose tracking software enables the rapid collection of large amounts of dose data, thereby contributing to the dose optimization process and making it possible to update DRLs at more frequent intervals. These types of software allow for the efficient monitoring of health center dose data and prompt initiation of the optimization process when dose values exceed reference DRLs.[5]

Several recent studies have shown that DRLs established based on patient size and clinical indications could significantly con-

tribute to dose optimization.[7,13,26-28] Clinical indication-based DRLs are established using dose data obtained from examinations performed for various indications and requiring different image quality. In the study conducted by Aberle et al.[13], the DRL for an abdomen CT protocol performed for the exclusion of kidney stones was found to be 45% lower than the DRLs for abdomen CT protocols for other indications. It has been noted that DRLs are strongly dependent on clinical indications.[13] The ICRP also emphasized the significance of clinical indications-based DRLs.[3] The European Society of Radiology initiated the European Study on Clinical Diagnostic Reference Levels for X-ray Medical Imaging project for the establishment of clinical indication-based DRLs.[26] Due to the lack of protocols created for different clinical indications in our institution, clinical indication-based DRLs could not be established in our study. To conduct future DRL studies based on clinical indications and to improve the success of the dose optimization process, it is necessary to develop protocols according to different clinical indications.

Klosterkemper et al.[27] showed significant variations in radiation doses based on patient sizes. Kanal et al.[7] established the NDRLs for the 10 most common CT examinations for adults in the United States, based on patient size (achievable dose according to water-equivalent diameter). In their prospective multicenter study, Brat et al.[28] established LDRLs for chest and abdomen CT examinations based on clinical indications and body mass index (BMI) class. Different dose levels were identified in different BMI classes, and particularly high variations were observed in doses for patients with a BMI ≥25.[28] DRLs could not be established based on patient sizes in our study due to the lack of weight information for the patients in our dataset. DRLs that are established based on patient sizes can contribute to the optimization of protocols and the prevention of unnecessary radiation exposure by reducing dose variations.

CTDIvol and DLP are indirect measurements of patient radiation dose. It is well-known that patient size affects radiation dose. To improve the accuracy of dose exposure measurements, patient size should also be considered.[7,29] Size-specific dose estimate (SSDE) is a method that recalculates the CTDIvol based on patient size, providing a more accurate prediction of the patient's radiation dose.[30] The ED is a quantity that represents the stochastic risk caused by radiation.[31]

In our study, the SSDE and ED, which were automatically calculated by the Teamplay software, were obtained from dose reports. However, SSDE values could not be verified due to the unavailability of patient height and weight data. There is currently no study validating the calculations used by the Teamplay software for determining ED. Therefore, SSDE and ED were not included in our study. The ICRP recommends establishing DRLs for pediatric examinations based on patient weight.[3] Hence, pediatric examinations were not included in the study. By obtaining patient size information, pediatric DRLs could be established, and dose metrics such as SSDE could be included in future studies.

To reduce radiation dose in coronary CTA, prospective gating mode (PGM) has been developed as an alternative to retrospective gating mode (RGM) for patients with stable heart rates. PGM has shown a significant dose reduction compared to RGM.[32] Therefore, in current coronary CTA DRL studies, separate DRLs for each mode have been established.[33] In our study, information regarding the mode in which coronary CTAs were performed was unavailable; thus, DRLs based on imaging mode could not be established. LDLRs for coronary CTA were compared with mixed-mode DRLs from other countries (Table 4). This issue could be resolved by labeling the different coronary CTA modes with separate protocol names.

Diagnostic quality should be preserved in parallel with doses being reduced during the optimization process. DRL studies not only establish radiation dose values that should not be exceeded but also demonstrate the minimum dose levels that can provide diagnostic quality. In cases where local dose levels are below the 25th percentile, image quality should be assessed.[3] One of the limitations of our study is that the image quality was not evaluated.

Despite its limitations, our study emphasizes the need for the re-establishment of NDRLs. Individual healthcare facilities should initiate the optimization process by monitoring their dose data.

In conclusion, LDRLs for CT were established from substantial dose data using dose-tracking software. There is a need to update the NDRLs for CT protocols in Türkiye. Until new NDRLs are established, local institutions in Türkiye can initiate the optimization process by comparing their dose distributions to the LDRLs established in our study. The LDRLs were comparable to or lower than NDRLs and DRLs set in other countries for most protocols. Automated dose-tracking software can play an important role in establishing DRLs by facilitating the collection and analysis of large datasets. The establishment and use of DRLs, as well as radiation dose optimization, can be achieved through the collaborative and coordinated efforts of radiologists, medical physicists, radiographers, and radiological safety officers. We hope that our study can contribute to radiation dose optimization efforts in Türkiye.

Conflict of interest disclosure

The authors declared no conflicts of interest.

References

1. Radiation UNSCotEoA. Sources, Effects and Risks of Ionizing Radiation, United Nations Scientific Committee on the Effects of Atomic Radiation (UNSCEAR) 2020/2021 Report, Volume I. United Nations; 2022. [CrossRef]

2. No authors listed. The 2007 Recommendations of the International Commission on Radiological Protection. ICRP publication 103. *Ann ICRP*. 2007;37(2-4):1-332. [CrossRef]

3. Vañó E, Miller DL, Martin CJ, et al. ICRP Publication 135: diagnostic reference levels in medical imaging. *Ann ICRP*. 2017;46(1):1-144. [CrossRef]

4. Brink JA, Miller DL. U.S. National diagnostic reference levels: closing the gap. *Radiology*. 2015;277(1):3-6. [CrossRef]

5. Loose RW, Vano E, Mildenberger P, et al. Radiation dose management systems-requirements and recommendations for users from the ESR EuroSafe imaging initiative. *Eur Radiol*. 2021;31(4):2106-2114. [CrossRef]

6. Ataç GK, Parmaksız A, İnal T, et al. Patient doses from CT examinations in Turkey. *Diagn Interv Radiol*. 2015;21(5):428-434. [CrossRef]

7. Kanal KM, Butler PF, Sengupta D, Bhargavan-Chatfield M, Coombs LP, Morin RL. U.S. diagnostic reference levels and achievable doses for 10 adult CT examinations. *Radiology*. 2017;284(1):120-133. [CrossRef]

8. Kanda R, Akahane M, Koba Y, et al. Developing diagnostic reference levels in Japan. *Jpn J Radiol*. 2021;39(4):307-314. [CrossRef]

9. Wardlaw GM, Martel N. Sci-Thur PM – colourful interactions: highlights 07: Canadian computed tomography survey: national diagnostic reference levels. *Med Phys*. 2016;43(8Part3):4932-4933. [CrossRef]

10. Agency UHS. UKHSA-RCE-1: doses from computed tomography (CT) exams in the UK: 2019 review. UK Health Security Agency; 2022. [CrossRef]

11. Tsapaki V, Damilakis J, Paulo G, et al. CT diagnostic reference levels based on clinical indications: results of a large-scale European survey. *Eur Radiol*. 2021;31(7):4459-4469. [CrossRef]

12. Schegerer A, Loose R, Heuser LJ, Brix G. Diagnostic reference levels for diagnostic and interventional X-ray procedures in Germany: update and handling. *Rofo*. 2019;191(8):739-751. [CrossRef]

13. Aberle C, Ryckx N, Treier R, Schindera S. Update of national diagnostic reference levels for adult CT in Switzerland and assessment of radiation dose reduction since 2010. *Eur Radiol*. 2020;30(3):1690-1700. [CrossRef]

14. Nam S, Park H, Kwon S, et al. Updated national diagnostic reference levels and achievable doses for ct protocols: a national survey of Korean hospitals. *Tomography*. 2022;8(5):2450-2459. [CrossRef]

15. No authors listed. Radiological protection and safety in medicine. A report of the International Commission on Radiological Protection. *Ann ICRP*. 1996;26(2):1-47. [CrossRef]

16. Hart D, Hillier MC, Wall BF. National reference doses for common radiographic, fluoroscopic and dental X-ray examinations in the UK. *Br J Radiol*. 2009;82(973):1-12. [CrossRef]

17. Nassiri MA, Rouleau M, Després P. CT dose reduction: approaches, strategies and results from a province-wide program in Quebec. *J Radiol Prot*. 2016;36(2):346-362. [CrossRef]

18. Ning P, Zhu S, Shi D, Guo Y, Sun M. X-ray dose reduction in abdominal computed tomography using advanced iterative reconstruction algorithms. *PLoS One*. 2014;9(3):e92568. [CrossRef]

19. Power SP, Moloney F, Twomey M, James K, O'Connor OJ, Maher MM. Computed tomography and patient risk: facts, perceptions and uncertainties. *World J Radiol*. 2016;8(12):902-915. [CrossRef]

20. Héliou R, Normandeau L, Beaudoin G. Towards dose reduction in CT: patient radiation dose assessment for CT examinations at university health center in Canada and comparison with national diagnostic reference levels. *Radiat Prot Dosimetry*. 2012;148(2):202-210. [CrossRef]

21. Protection NCoR, Levels MSC-oDR, Achievable Doses, States RLiMIRfAitU, Protection NCoR, Measurements. Reference Levels and Achievable Doses in Medical and Dental Imaging: Recommendations for the United States. National Council on Radiation Protection and Measurements; 2012. [CrossRef]

22. Sharma R, Sharma SD, Pawar S, Chaubey A, Kantharia S, Babu DA. Radiation dose to patients from X-ray radiographic examinations using computed radiography imaging system. *J Med Phys*. 2015;40(1):29-37. [CrossRef]

23. Hojreh A, Weber M, Homolka P. Effect of staff training on radiation dose in pediatric CT. *Eur J Radiol*. 2015;84(8):1574-1578. [CrossRef]

24. Paolicchi F, Faggioni L, Bastiani L, Molinaro S, Caramella D, Bartolozzi C. Real practice radiation dose and dosimetric impact

of radiological staff training in body CT examinations. *Insights Imaging*. 2013;4(2):239-244. [CrossRef]

25. Wang KC, Patel JB, Vyas B, et al. Use of radiology procedure codes in health care: the need for standardization and structure. *Radiographics*. 2017;37(4):1099-1110. [CrossRef]

26. Paulo G, Damilakis J, Tsapaki V, et al. Diagnostic reference levels based on clinical indications in computed tomography: a literature review. *Insights Imaging*. 2020;11(1):96. [CrossRef]

27. Klosterkemper Y, Appel E, Thomas C, et al. Tailoring CT dose to patient size: implementation of the updated 2017 ACR size-specific diagnostic reference levels. *Acad Radiol*. 2018;25(12):1624-1631. [CrossRef]

28. Brat H, Zanca F, Montandon S, et al. Local clinical diagnostic reference levels for chest and abdomen CT examinations in adults as a function of body mass index and clinical indication: a prospective multicenter study. *Eur Radiol*. 2019;29(12):6794-6804. [CrossRef]

29. Waszczuk Ł A, Guziński M, Czarnecka A, Sąsiadek MJ. Size-specific dose estimates for evaluation of individual patient dose in CT protocol for renal colic. *AJR Am J Roentgenol*. 2015;205(1):100-105. [CrossRef]

30. Brady SL, Kaufman RA. Investigation of American Association of Physicists in Medicine Report 204 size-specific dose estimates for pediatric CT implementation. *Radiology*. 2012;265(3):832-840. [CrossRef]

31. Fisher DR, Fahey FH. Appropriate use of effective dose in radiation protection and risk assessment. *Health Phys*. 2017;113(2):102-109. [CrossRef]

32. Sun Z. Multislice CT angiography in cardiac imaging: prospective ECG-gating or retrospective ECG-gating? *Biomed Imaging Interv J*. 2010;6(1):e4. [CrossRef]

33. Alhailiy AB, Brennan PC, McEntee MF, Kench PL, Ryan EA. Diagnostic reference levels in cardiac computed tomography angiography: a systematic review. *Radiat Prot Dosimetry*. 2018;178(1):63-72. [CrossRef]