Original Investigation

Automated Assessment of T2-Weighted MRI to Differentiate Malignant and Benign Primary Solid Liver Lesions in Noncirrhotic Livers Using Radiomics

Martijn P.A. Starmans, PhD, Razvan L. Miclea, MD, PhD, Valerie Vilgrain, MD, PhD, Maxime Ronot, MD, PhD, Yvonne Purcell, MD, Jef Verbeek, MD, PhD, Wiro J. Niessen, PhD, Jan N.M. Ijzermans, MD, PhD, Rob A. de Man, MD, PhD, Michael Doukas, MD, PhD, Stefan Klein, PhD[1], Maarten G. Thomeer, MD, PhD[1]

Rationale and Objectives: Distinguishing malignant from benign liver lesions based on magnetic resonance imaging (MRI) is an important but often challenging task, especially in noncirrhotic livers. We developed and externally validated a radiomics model to quantitatively assess T2-weighted MRI to distinguish the most common malignant and benign primary solid liver lesions in noncirrhotic livers.

Materials and Methods: Data sets were retrospectively collected from three tertiary referral centers (A, B, and C) between 2002 and 2018. Patients with malignant (hepatocellular carcinoma and intrahepatic cholangiocarcinoma) and benign (hepatocellular adenoma and focal nodular hyperplasia) lesions were included. A radiomics model based on T2-weighted MRI was developed in data set A using a combination of machine learning approaches. The model was internally evaluated on data set A through cross-validation, externally validated on data sets B and C, and compared to visual scoring of two experienced abdominal radiologists on data set C.

Results: The overall data set included 486 patients (A: 187, B: 98, and C: 201). The radiomics model had a mean area under the curve (AUC) of 0.78 upon internal validation on data set A and a similar AUC in external validation (B: 0.74 and C: 0.76). In data set C, the two radiologists showed moderate agreement (Cohen's κ: 0.61) and achieved AUCs of 0.86 and 0.82.

Conclusion: Our T2-weighted MRI radiomics model shows potential for distinguishing malignant from benign primary solid liver lesions. External validation indicated that the model is generalizable despite substantial MRI acquisition protocol differences. Pending further optimization and generalization, this model may aid radiologists in improving the diagnostic workup of patients with liver lesions.

Key Words: Hepatocellular carcinoma; Liver cancer; Machine learning; Magnetic resonance imaging.

© 2023 The Association of University Radiologists. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Acad Radiol xxxx; xx:xxx–xxx

From the Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands (M.P.A.S., W.J.N., S.K., M.G.T.); Department of Radiology and Nuclear Medicine, Maastricht UMC+, Maastricht, the Netherlands (R.L.M.); Université de Paris, INSERM U 1149, CRI, Paris, France (V.V., M.R.); Département de Radiologie, Hôpital Beaujon, APHP.Nord, Clichy, France (V.V., M.R.); Department of Radiology, Hôpital Fondation Rothschild, Paris, France (Y.P.); Department of Gastroenterology and Hepatology, University Hospitals Leuven, Leuven, Belgium (J.V.); Department of Gastroenterology and Hepatology, Maastricht UMC+, Maastricht, the Netherlands (J.V.); Faculty of Applied Sciences, Delft University of Technology, the Netherlands (W.J.N.); Department of Surgery, Erasmus MC, Rotterdam, the Netherlands (J.N.M.I.); Department of Gastroenterology & Hepatology, Erasmus MC, Rotterdam, the Netherlands (R.A.d.M.); Department of Pathology, Erasmus MC, Rotterdam, the Netherlands (M.D.). Received March 22, 2023; revised July 6, 2023; accepted July 25, 2023.
Address correspondence to: M.P.A.S. e-mail: m.starmans@erasmusmc.nl

[1] Equal contributions.

© 2023 The Association of University Radiologists. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
https://doi.org/10.1016/j.acra.2023.07.024

INTRODUCTION

Liver cancer is the seventh most commonly diagnosed cancer and the third most common cause of cancer death worldwide, with approximately 906,000 estimated new cases and 830,000 deaths in 2020 (1). One of the most important tasks in routine clinical practice is to distinguish malignant from benign primary solid liver lesions, as this differentiation influences treatment planning. The final diagnosis is often based on biopsy (2,3), but the first assessment is typically performed by radiologists on magnetic resonance imaging (MRI). In patients with underlying chronic liver disease (e.g., cirrhosis), assessment is relatively easy, as the a priori chance of a lesion being hepatocellular carcinoma (HCC) is by far the largest (4). Furthermore, the Liver Imaging Reporting and Data System (LI-RADS) system (5), specifically designed to aid radiologists in classifying liver lesions in cirrhotic livers, has been shown to be quite effective. In noncirrhotic livers however, diagnosis is more challenging due to the wide variety of phenotypes (6) and lack of a clear MRI assessment consensus (7). Specifically for HCC in noncirrhotic livers, there has been an increase in prevalence, requiring further investigation and diagnostic aids as the phenomenon remains complex (8).

In recent years, radiomics—the use of large numbers of quantitative medical imaging features to predict clinical outcomes—has been successfully used to create diagnostic aids in various clinical areas (9). In liver cancer, the use of radiomics has mostly focused on computed tomography (10). Regarding MRI in liver cancer, radiomics has been used to classify focal liver lesions (11–15) and hepatic lesions (16), and as LI-RADS surrogate (5,17). However, the characterization of liver lesions based on radiomics is still at an early stage (18). Major challenges include the lack of large multicenter cohorts, particularly for external validation, and image acquisition heterogeneity.

Diagnosis on MRI is commonly based on a variety of sequencies, for example, T2-weighted MRI, T1-weighted MRI with or without (liver-specific) contrast agents, and diffusion-weighted imaging (2,3). The presence of hypervascularity, wash-out, liquid, and intralesional fat on these sequences can commonly be made with high accuracy by the radiologist as illustrated by LI-RADS. However, while aided by guidelines (2,3), the radiologist's interpretation of T2-weighted MRI remains challenging, qualitative, and observer dependent. For example, while some heterogeneity patterns are associated with malignancy (e.g., vast chaotic heterogeneity (19) or mosaic (20)), others (e.g., atoll or crescent sign (21,22)) make the diagnosis of liver cancer less probable. In this study, we propose to evaluate radiomics as a diagnostic aid to objectively estimate the probability of solid liver lesion malignancy based on T2-weighted MRI. This sequence is frequently used and widely available in the assessment of liver cancer, is relatively straightforward to perform, shows less heterogeneity, and is less invasive compared to sequences using contrast agents (2,3,5). Hence, a T2-weighted MRI–based radiomics model would be widely applicable, scalable, and generalizable.

The primary aim of this study was to evaluate radiomics in the quantitative assessment of T2-weighted MRI to distinguish malignant and benign primary solid liver lesions in noncirrhotic livers. As a pilot study, we focused on the most common lesion types, which concern more than 90% of all primary solid liver tumors: HCC, intrahepatic cholangiocarcinoma (iCCA), focal nodular hyperplasia (FNH), and hepatocellular adenoma (HCA). We externally validated the radiomics model in two multicenter cohorts and compared its performance to visual assessment by two experienced abdominal radiologists.

MATERIALS AND METHODS

Data Collection

Three data sets were collected retrospectively from three tertiary referral centers: all patients diagnosed at or referred to (A) the Erasmus University Medical Center, Rotterdam, the Netherlands, between 2002 and 2018 (publicly released (23)); (B) Maastricht UMC+, Maastricht, the Netherlands, between 2005 and 2018; and (C) Hôpital Beaujon, Paris, France, included in reverse chronological order starting in 2018, until a total of 201 patients were identified. Imaging data, age, sex, and phenotype were collected for each patient.

Inclusion criteria were as follows: HCC, iCCA, HCA, or FNH; pathologically proven phenotype, except "typical" FNH; and availability of a T2-weighted MRI scan. Exclusion criteria were as follows: maximum diameter ≤3 cm; underlying liver disease; or significant imaging artifacts.

Malignant lesions included HCC (75%–85% of primary liver cancers (6)) and iCCA (10%–15% of primary liver cancers) (6). Benign lesions included HCA (3–4 cases per 100,000 person-years in Europe and North America) and FNH (found in 0.8% of all adult autopsies) (6). The most common benign primary liver lesions, hemangiomas, and cysts were not included, as these are nonsolid and generally easy to diagnose on imaging (2,24). Only lesions with a pathologically proven phenotype were included to ensure an objective ground truth. Pathological analysis for each patient was performed locally in their admission hospital. Details of the pathological examination are provided in Supplementary Material 1. An exception was made for typical FNH (6), which are routinely diagnosed radiologically and not biopsied (25), as typical FNH imaging characteristics are 100% specific (2). Excluding these would have created a selection bias toward "atypical" FNH, meaning that no claims could be made regarding model performance in typical FNH.

Lesions with a maximum diameter ≤3 cm were excluded, since in noncirrhotic livers, these have a higher probability of being secondary lesions, hemangiomas, or cysts (24,26), which are generally easy to diagnose on imaging (2,24).

Specifically for secondary lesions, the diagnosis is further simplified by the patient's clinical background and history; for example, it is frequently already known whether a primary tumor is present at another location upon secondary lesion detection.

Patients with underlying liver disease due to alcohol, hepatitis, or vascular liver disease, such as fibrosis or cirrhosis, were excluded. Steatosis was not an exclusion criterion, as a degree of steatosis frequently occurs in the general population (27). Diagnosis of liver disease was based on clinical, pathological, and/or imaging findings. In patients with HCC, absence of cirrhosis was always confirmed by biopsy or resection.

When T2-weighted MRI with fat saturation was not available, regular T2-weighted MRI was used, similar to clinical practice. Images with significant artifacts (i.e., patient related or scanner related) and therefore not suitable for diagnostic purposes, as judged by an experienced radiologist (21 years of experience), were excluded. In patients with multiple lesions, only the largest was included.

Segmentation

Lesions were semiautomatically segmented using in-house software (12) by one of the three observers: a radiology resident and two experienced abdominal radiologists (RAD1 (M.G.T.): 21 years of experience and RAD2 (R.L.M): 8 years of experience). A subset of 60 lesions (data set B: 30 and data set C: 30) was segmented by two observers (RAD1 and RAD2) to assess the interobserver variability using the pairwise Dice Similarity Coefficient (DSC) (28).

Radiomics

An overview of the radiomics methodology is depicted in Figure 1. All radiomics steps were performed using the Workflow for Optimal Radiomics Classification (WORC) toolbox (29,30). The code for feature extraction and model creation is available as open source (31).

Before feature extraction, the images were normalized using z-scoring; no other preprocessing steps were applied. Instead of correcting for variations across centers, we hypothesized that training the model on heterogeneous, multicenter, representative data will facilitate generalization. For each lesion, the default set of 564 features from WORC quantifying intensity, shape, and texture were extracted from the T2-weighted MRI scan (29).

Construction of a radiomics model from the wide variety of available methods manually through a heuristic trial-and-error process has various disadvantages; for example, it is time consuming, not reproducible, does not guarantee an optimal solution, and has a high risk of overfitting. Hence, we instead used the WORC algorithm to automate and optimize this process (29). In WORC, decision model creation consists of several standardized components, for example, feature selection, resampling, and machine learning. For each component, a large collection of commonly used algorithms and their associated hyperparameters is included. For example, for the classification component, WORC includes eight different algorithms: support vector machine, random forest, logistic regression, linear and quadratic discriminant analysis, Gaussian Naïve Bayes, AdaBoost, and extreme gradient boosting (XGBoost). WORC exploits automated machine learning to compare 1000 different radiomics workflows (i.e., specific, randomly selected combinations of algorithms and hyperparameters) and optimize the combination that maximizes prediction performance on the training data set (29). The final model consists of an ensemble of the top 100 performing workflows by averaging their posterior probabilities.

Experimental Setup

First, to evaluate the predictive value of radiomics within a single center, internal validation was performed in data set A through a 100× stratified random-split cross-validation (32,33), see Supplementary Figure S1a. Second, to evaluate whether the model generalizes well to unseen data from other centers, two external validations were performed by training a model on data set A and testing it on the unseen data sets B and C, see Supplementary Figure S1b. Third, as clinicians frequently use age and sex in their decision-making, two additional models were externally validated based on (1) age and sex; and (2) age, sex, and radiomics features.

For both the internal and external validations, all model construction and optimization were performed within the training data set using an internal 5× random-split cross-validation in order to prevent overfitting on the test data set, see Supplementary Figure S1.

Performance of the Radiologists

To compare the models to visual assessment, the T2-weighted MRI scans were scored by two experienced abdominal radiologists (RAD1 and RAD2). They were blinded to the diagnosis and only had access to the T2-weighted MRI but were aware of the inclusion and exclusion criteria of the study. Classification of malignancy was made on a four-point scale to indicate the radiologists' certainty: 1 = benign, certain; 2 = benign, uncertain; 3 = malignant, uncertain; and 4 = malignant, certain. Additionally, the radiologists scored several characteristics used in the decision-making: the presence of (1) central scar (6); (2) liquid; (3) atoll sign (22); and (4) degree of heterogeneity (scale 1–4 similar to malignancy). As the radiologists were from centers A and B, scoring was carried out on data set C to prevent previous exposure to the data.

Model Insight

To gain insight into the radiomics model's decision-making, lesions were ranked based on the probability of a lesion being malignant as predicted by the model. Ranking was done as archetypal benign (ground truth benign, probability near 0%) - pitfall malignant (ground truth malignant, probability near 0%) - borderline (probability around 50%) - pitfall benign

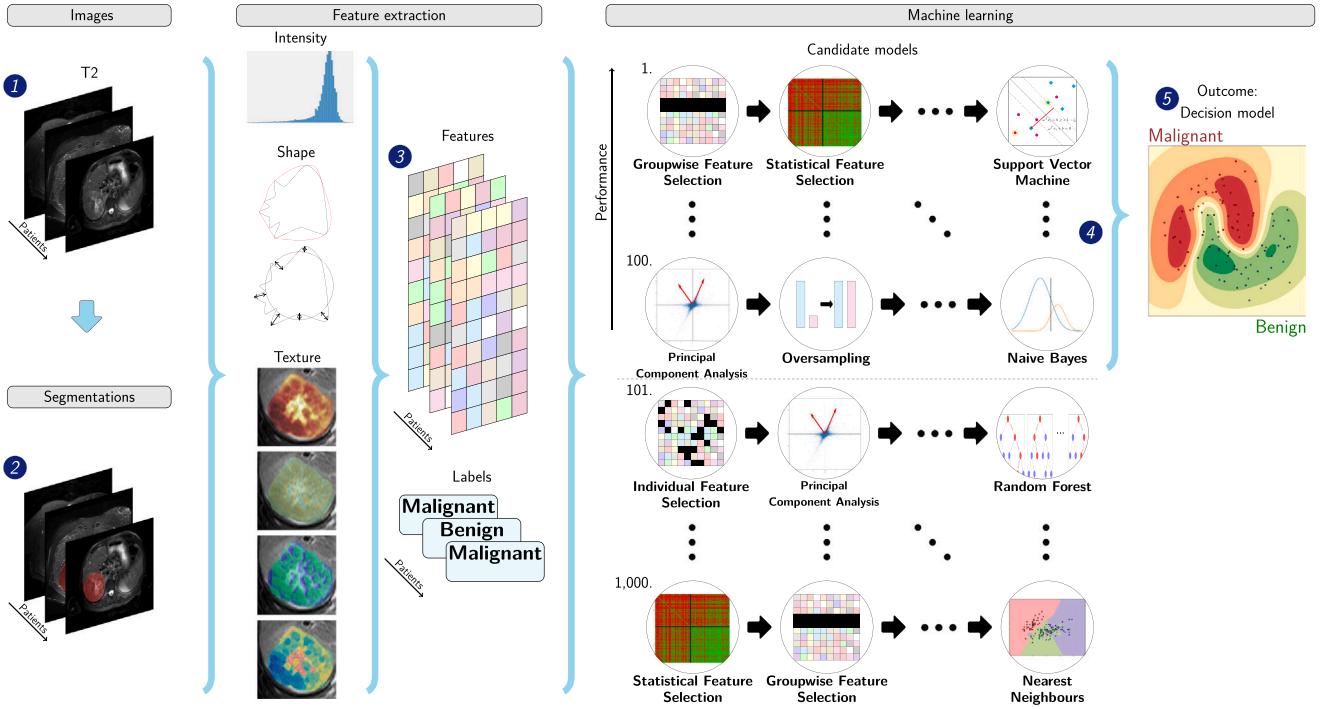

Figure 1. Schematic overview of the radiomics approach. Input to the algorithm are the T2-weighted magnetic resonance imaging scans (1) and the lesion segmentations (2). Processing steps include feature extraction (3) and the creation of a machine learning decision model (5) using an ensemble of the best 100 workflows from 1000 candidate workflows (4), where the workflows are different combinations of the different analysis steps (e.g., the classifier used). Adapted from Vos et al. (45): the images under (1), texture features, numbers at (3), and output at (4) have been modified with respect to the original figure.

(ground truth benign, probability near 100%) - archetypal malignant (ground truth malignant, probability near 100%). This was done using data set C to enable comparison with the radiologists.

Statistical Analysis

To evaluate the differences in clinical characteristics and radiomics features between the malignant and benign lesions, for each data set separately, univariate statistical testing was performed using a Mann-Whitney U test for continuous variables and a Chi-square test for categorical variables. P-values of the radiomics features were corrected using the Bonferroni correction (i.e., multiplying the p-values by the number of tests). To analyze the differences between data sets, the distributions of clinical characteristics were statistically compared using a Kruskal-Wallis test for continuous variables and a Chi-square test for discrete variables.

For all radiomics models, the area under the curve (AUC) of the receiver operating characteristic (ROC) curve, as well as accuracy, sensitivity, and specificity, were calculated. The positive class was defined as the malignant lesions. For the internally validated model, 95% confidence intervals of the performance metrics were constructed using the corrected resampled t-test, thereby taking into account that the samples in the cross-validation splits are not statistically independent (33). For the externally validated model, 95% confidence intervals were constructed using 1000× bootstrap resampling of the test data set and the standard method for normal distributions ((34) Table 6, method 1). ROC confidence bands were constructed using fixed-width bands (35).

For binary scores, the agreement between radiologists was evaluated using Cohen's κ (36). For ordinal scores, that is, degree of heterogeneity and malignancy, the correlation was evaluated using Spearman's ρ (37). The AUCs of the radiomics model and the radiologists were compared using the DeLong test (38), and confusion matrices were used to analyze agreement.

For all statistical tests, p-values < 0.05 were considered statistically significant. As the performances of the radiomics model and radiologists were not known beforehand, no a priori sample size calculation could be performed. Instead, sample size is taken into account naturally in the used confidence interval computations, and power tests were performed for the DeLong test (38).

All statistical analyses were performed using Python 3.7.6 (the Python Software Foundation).

RESULTS

Data Sets

In total, 486 patients were included (A: 187, B: 98, C: 201). The clinical and imaging characteristics are reported in

TABLE 1. Clinical and Imaging Characteristics of the Data Sets

Data Set	A: Erasmus MC (*n* = 187)			B: Maastricht UMC+[a] (*n* = 98)			C: Beaujon APHP (*n* = 201)			p
	Benign	Malignant	p	Benign	Malignant	p	Benign	Malignant	p	
Patients	93	94		55	43		117	84		
Age (years)[a]	37 (30–46)	62 (25–70)	**10^{-19}**	38 (31–45)	64 (60–71)	**10^{-14}**	38 (31–45)	63 (53–68)	**10^{-20}**	0.69
Sex			**10^{-12}**			**10^{-6}**			**10^{-17}**	0.22
Male	4	48		3	20		11	55		
Female	89	46		52	23		106	29		
Phenotype										**0.003**
HCC		81			28			47		
iCCA		13			15			37		
HCA	48			26			65			
FNH	45			29			52			
Volume (cl)[a,†]	11 (4–27)	17 (5–50)	>1.0	11 (4–24)	21 (6–59)	>1.0	8 (3–19)	9 (3–33)	>1.0	10^{-5}
Imaging										
Magnetic field strength			0.10			0.45			**0.003**	**0.002**
1.0 Tesla	1	4		2	4		1	3		
1.5 Tesla	76	82		48	39		74	68		
3.0 Tesla	16	8		5	0		42	13		
Scanner manufacturer			**10^{-8}**			**0.03**			0.77	**10^{-15}**
Siemens	13	32		21	7		23	17		
Philips	16	38		34	36		62	40		
GE	64	24		0	0		30	24		
Toshiba	0	0		0	0		2	3		
Slice thickness (mm)[b]	6.0–8.0	6.0–7.0	0.12	5.0–6.0	5.0–5.0	0.08	5.0–6.0	5.0–6.0	0.41	**10^{-32}**
Pixel spacing (mm)[b]	0.72–0.94	0.73–1.19	**0.005**	0.77–1.38	0.77–0.99	0.13	0.74–1.0	0.75–1.07	0.13	0.07
Repetition time (ms)[b]	1348–8571	1218–4844	**0.001**	1100–2805	1600–2961	**0.007**	1200–3884	1512–6058	0.14	**0.006**
Echo time (ms)[b]	89–100	80–100	**10^{-3}**	80–112	80–90	**0.04**	80–120	80–103	0.13	0.62
Flip angle (degree)[b]	90–150	90–150	0.47	90–141	90–90	**0.01**	90–140	90–134	0.33	**0.07**
Fat saturation (yes/no)	72/21	59 / 35	**0.04**	35/20	39/4	**0.004**	98/19	59/25	**0.03**	**10^{-18}**

FNH, focal nodular hyperplasia; GE, general electric; HCA, hepatocellular adenoma; HCC, hepatocellular carcinoma; iCCA, intrahepatic cholangiocarcinoma; Max, maximum.

The number of patients (n) in each data set is indicated in the column header. Per data set, the statistical significance of the difference between the malignant and benign lesions was assessed using a Mann-Whitney U test for continuous variables and a Chi-square test for discrete variables. The statistical significance of the difference between data sets was assessed using a Kruskal-Wallis test for continuous variables and a Chi-square test for discrete variables. Statistically significant p-values are displayed in bold.

[a] Median (quartile 1–quartile 3).

[b] Quartile 1–quartile 3.

[†] p-value after Bonferroni correction.

TABLE 2. Performance of the Radiomics Model and the Radiologists in the Three Data Sets (A, B, and C)

Evaluation	Internal Cross-validation	External Validation		Radiologist 1	Radiologist 2
Train set	A[a]	A	A	-	-
Test set	A[a]	B	C	C	C
AUC	0.80 (0.74, 0.85)	0.78 (0.69, 0.88)	0.75 (0.67, 0.82)	0.86	0.83
Accuracy	0.72 (0.65, 0.78)	0.71 (0.63, 0.80)	0.70 (0.64, 0.77)	0.80	0.77
Sensitivity	0.69 (0.60, 0.78)	0.84 (0.72, 0.96)	0.80 (0.71, 0.89)	0.88	0.87
Specificity	0.74 (0.64, 0.84)	0.62 (0.50, 0.73)	0.64 (0.56, 0.72)	0.74	0.69

AUC, area under the receiver operating characteristic curve.

For the radiomics model, the mean (internal cross-validation) or point estimate (external validation) and 95% confidence intervals are reported.

[a] Training and testing within a single data set were done through a 100× random-split cross-validation.

Table 1. As all centers serve as tertiary referral centers, the data sets originated from 159 different scanners (A: 52, B: 21, and C: 86), resulting in substantial MRI acquisition protocol heterogeneity. Statistically significant differences were found between data sets A, B, and C for magnetic field strength (Kruskal-Wallis: $p = 0.001$), manufacturer ($p = 10^{-4}$), slice thickness ($p = 10^{-32}$), repetition time ($p = 0.006$), flip angle ($p = 0.05$), and use of fat saturation ($p = 10^{-17}$).

On the subset that was segmented by two observers, the mean DSC indicated a high average overlap, but the standard deviation also indicated large discrepancies (B: 0.80 ± 0.21; C: 0.81 ± 0.11).

Radiomics

The results of the radiomics model are presented in Table 2. The internal validation on data set A had a mean AUC of 0.80; the two external validations yielded a similar performance (B: 0.75; C: 0.78). The ROC curves (Fig 2) illustrate that the model trained on data set A performed similarly in each of the three centers. The accuracy per phenotype is presented in Table 3. The radiomics model had a similar accuracy in HCC (0.80) and iCCA (0.78), while the performance in FNH (0.76) was better than in HCA (0.54).

The age-and-sex-only model had a high AUC in both the internal validation (A: 0.94) and the two external validations (B: 0.93 and C: 0.92). Combining age, sex, and the radiomics features yielded an improvement (A: 0.94, B: 0.98, and C: 0.98), although not statistically significant.

Comparison with Radiologists

The performance of the two experienced abdominal radiologists in classifying data set C is presented in Tables 2 and 3. The ROC curves (Fig 2c) were mostly just above the 95% confidence interval of the radiomics model. The AUC of RAD1 (0.86) was statistically significantly better than the radiomics model (DeLong: $p < 0.001$): the differences in AUC between RAD2 (0.83) and the radiomics model and between the two radiologists were not statistically significant.

Confusion matrices of the predictions on data set C are depicted in Figure 3. The agreement between the radiologists in classifying the lesions as malignant or benign was moderate (Cohen's κ (36): 0.61): the two radiologists agreed in 160/201 patients (80%). The agreement between the two

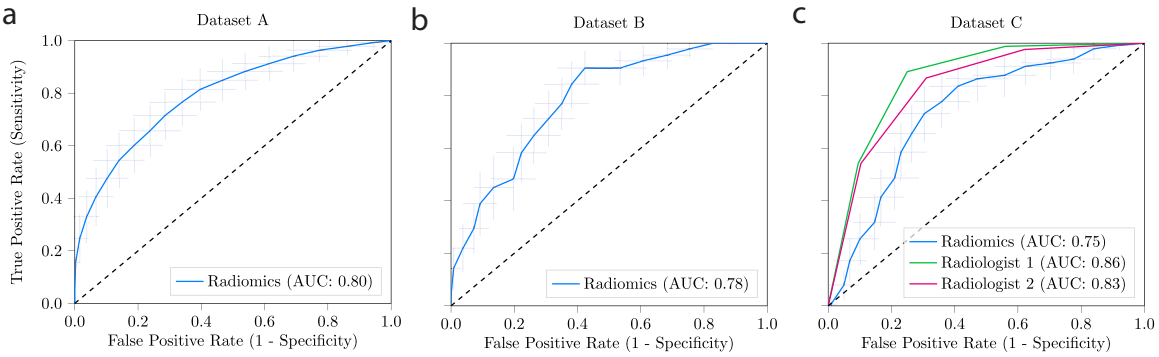

Figure 2. Receiver operating characteristic curves of the radiomics model and radiologists. For the radiomics model, the curves present the model internally validated on data set A (**a**); and trained on data set A, externally validated on data set B (**b**) and data set C (**c**). The performance of scoring by the two experienced abdominal radiologists on data set C is also depicted in (**c**). For the radiomics model, the crosses identify the 95% confidence intervals of the 100× random-split cross-validation (**a**) or 1000× bootstrap resampling (**b** and **c**); the bold curves are fit through the means. AUC, area under the receiver operating characteristic curve.

TABLE 3. Accuracy Per Phenotype of the Radiologists and the Radiomics Model in the External Validation on Data Set C

Accuracy	Radiomics	Radiologist 1	Radiologist 2
Train data set	A	-	-
Test data set	C	C	C
HCC (47)	0.80	0.85	0.83
iCCA (37)	0.78	0.95	0.92
HCA (65)	0.54	0.69	0.62
FNH (52)	0.76	0.82	0.78

FNH, focal nodular hyperplasia; HCA, hepatocellular adenoma; HCC, hepatocellular carcinoma; iCCA, intrahepatic cholangiocarcinoma.

The accuracy per phenotype represents the percentage of the lesions with that specific phenotype being correctly classified as malignant or benign. The number of lesions per phenotype in data set C is given between brackets in the first column.

radiologists and the radiomics model was moderate (RAD1 κ: 0.55) and weak (RAD2 κ: 0.45), as reflected by the confusion matrices. For the other characteristics scored by the two radiologists, the agreement was weak for the presence of a scar (κ: 0.41) and liquid (κ: 0.52) and strong for the presence of the atoll sign (κ: 0.80); the correlation was strong for heterogeneity (Spearman's ρ (37): 0.70) and malignancy (ρ: 0.70). The radiomics model correctly predicted 2/23 patients that both radiologists scored incorrectly and 23/62 patients that at least one radiologist scored incorrectly.

Model Insight

In data set A, on which the radiomics model was developed, 45 radiomics features showed statistically significant differences between the malignant and benign lesions, with p-values after Bonferroni correction ranging from 10^{-10} to 0.049. These included four shape features (volume was not significant), one orientation feature, and 40 texture features. Statistically significant differences were found for 49 radiomics features in data set B and 10 in data set C. Only four radiomics features (all texture features) showed statistically significant differences in all three data sets. A list of these features and their p-values can be found in Supplementary Table S1. The differences in volume between the three data sets were statistically significant ($p = 10^{-5}$).

Examples of lesions from data set C ranked as archetypal, borderline, or pitfall by the radiomics model are shown in Figure 4. Visual inspection of the T2-weighted MRI scans of the archetypal or pitfall lesions showed a relation with heterogeneity (archetypal malignant: heterogeneous; archetypal benign: homogeneous), area and volume (archetypal malignant: generally high maximum axial area and high volume), and irregularity of shape on two-dimensional axial slices (archetypal malignant lesions: irregular; archetypal benign: compact). Pitfall lesions showed the opposite, for example, pitfall benign: heterogeneous. Borderline lesions, that is, with an almost equal predicted chance of being malignant or benign, were mostly of medium size and medium heterogeneity.

The predictions by the radiomics model on data set C were compared to the characteristic scores of RAD1, who had the highest performance. The correlation between the probability of malignancy as predicted by the radiomics model and heterogeneity as scored by RAD1 was moderate (ρ: 0.59). RAD1 performed well when lesions had an apparent atoll sign: from the 19 lesions which RAD1 scored as having an atoll sign and therefore classified as benign, 17 were indeed benign, and two were malignant. On the contrary, the radiomics model only classified 11 of these lesions correctly, but these included the two malignant lesions misclassified by RAD1.

DISCUSSION

In this pilot, we developed a radiomics model to quantitatively assess T2-weighted MRI to distinguish between

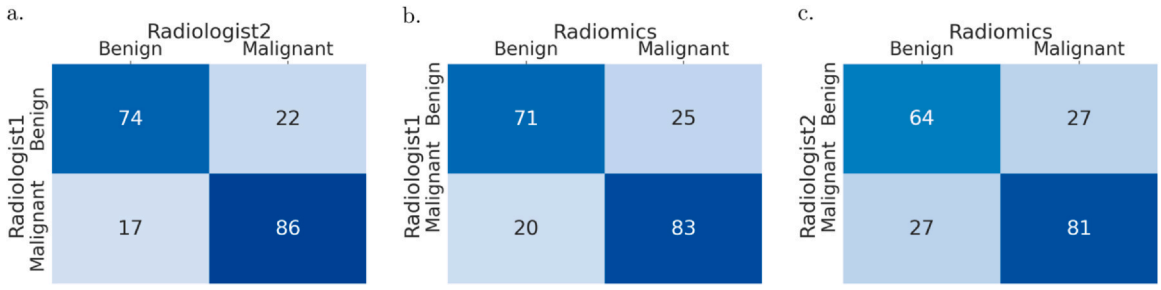

Figure 3. Confusion matrices of the predictions by the radiomics model and the two radiologists. The darker the background, the higher the agreement.

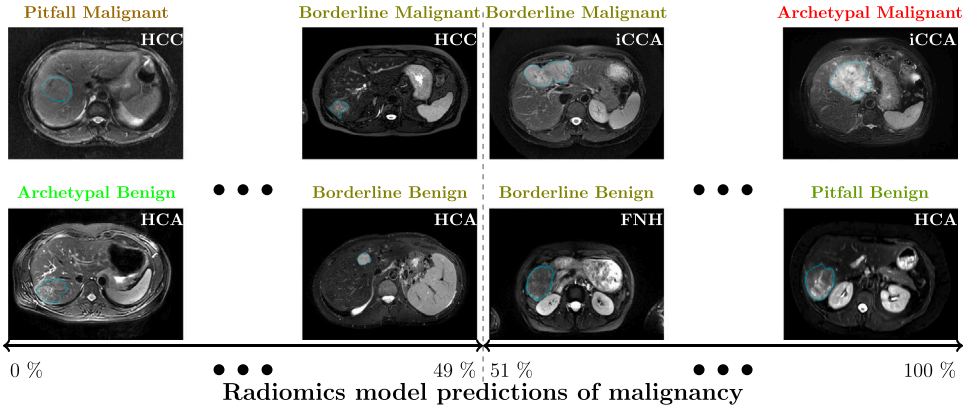

Figure 4. Examples of liver lesions on T2-weighted MRI. From left to right, examples of lesions considered by the radiomics model as archetypal (i.e., predicted probability close to extremes and correct), pitfall (i.e., predicted probability close to extremes and incorrect), and borderline (i.e., predicted probability close to border of 50%). FNH, focal nodular hyperplasia; HCA, hepatocellular adenoma; HCC, hepatocellular carcinoma; iCCA, intrahepatic cholangiocarcinoma.

malignant and benign primary solid liver lesions in noncirrhotic livers. We showed that our radiomics model can distinguish between these lesions, both in an internal cross-validation and two external validations, but that caution is warranted concerning clinical utility at this stage.

Currently, visual T2-weighted MRI interpretation is challenging, qualitative, and observer dependent, as supported by our results on the difference in interpretation by two experienced readers. Radiomics may help to overcome these deficits. The moderate correlation between the probability of malignancy by the radiomics model and the first radiologist and the differences in typical characteristics apparently used by the radiomics model (e.g., irregularity of shape) and the radiologist (e.g., atoll sign) indicate the potential complementary value of radiomics. Further research should focus on how radiologists can optimally use radiomics assessment of T2-weighted MRI in their clinical workflow and how to optimally combine imaging with clinical factors such as age and sex. Radiomics may be especially useful when there is no consensus between radiologists or in pitfalls for radiologists. Future work should also include improving the explainability and interpretability of our radiomics model, for example, by using techniques such as SHAP and LIME (39) to quantify feature importance in the model's decision-making.

While in recent years, several studies have evaluated radiomics for liver lesion classification (11–17,40), it is still at an early stage (18) and contains common radiomics vulnerabilities (9,41–43). With respect to existing literature, our study addresses several important issues. First, to ensure a high level of evidence, our model was externally validated in two multicenter cohorts from different countries. Second, we used routinely acquired T2-weighted MRI without strict protocol requirements. Despite substantial acquisition protocol heterogeneity, in line with our hypothesis, our method showed similar performance in internal and external validation, increasing the chance that the reported performance can be reproduced in a routine clinical setting. Third, our method uses automated machine learning to determine the optimal radiomics pipeline from a large number of methods. This facilitates reproducibility, automatically compares a large number of methods to optimize performance, and prevents overly optimistic performance estimates.

Age and sex are strong predictors for distinguishing malignant from benign liver lesions (1,24). In our study, in line with international findings, the majority of benign lesions were found in (young) females, while the majority of malignant lesions were found in older patients (1,24). The models based on age and sex used an age threshold set at 49 years, but in data set C, 19/114 (17%) lesions in patients aged < 49 years were malignant. Although this threshold yielded a high overall performance, it would lead to missing all malignant lesions in younger patients, the group which benefits the most from accurate and timely diagnosis. Simply classifying all lesions in patients aged < 49 years as benign, regardless of imaging information, is unacceptable and cannot be applied to the general population. Moreover, the relation in our database may be too strong due to our inclusion and exclusion criteria and thereby not representative of clinical practice. The T2-weighted MRI radiomics model does not use population-based information but predicts the lesion malignancy probability based purely on imaging appearance. Radiomics could be especially useful to avoid missing malignant lesions in young males and to detect benign lesions in older females. Future research should include optimally combining imaging, age, sex, and other clinical factors in the clinical workflow where radiologists typically have access to this information. For example, the radiologists can take these factors and the outcomes of a purely imaging-based radiomics model, which is less biased toward population-based information and more interpretable, separately into account

in decision-making. Alternatively, an integrated radiomics model optimally combining clinical factors and imaging can be created, and the outcome presented to the radiologists. While such an integrated model may overall perform better than imaging only, it is more susceptible to biases as also observed in our study and may be less interpretable.

Our study has several limitations. First, while the inclusion and exclusion criteria were set to maximize relevance to clinical decision-making of this pilot, they limit applicability. Future research should focus on loosening these criteria, for example, including smaller lesions, secondary lesions, non-solid lesions, and patients with liver disease. We expect that loosening these criteria will also give a more representative distribution of clinical practice in terms of age and sex, thus allowing research into optimally combining imaging, age, and sex. Second, the current approach requires semiautomatic segmentations. While accurate, this is time consuming and subject to some observer variability, hindering transition to clinical practice. Future research should therefore include automatic segmentation of the tumors, for example, using deep learning (44). We do however not believe that inter-observer variability substantially affected the results, as the radiomics model showed similar performance in the internal and external validations despite training and testing on segmentations from different observers. Depending on tumor size, the complete radiomics model takes a few seconds to maximally a minute to execute.

Future research should focus on extending this pilot to phenotyping to further aid clinical decision-making. Future research should also include evaluation of our model in the intended setting: radiomics to quantitatively assess T2-weighted MRI combined with visual assessment of other sequences. By isolating the T2-weighted MRI, we could evaluate the value of quantitative assessment of this sequence in detail, independently of other sequences. In real life, radiologists use multiple sequences in their assessment (2,3), suggesting that they contain additional information (13). As a result of the lack of standardized protocols in the literature, the main additional challenges facing multisequence radiomics are additional heterogeneity and missing data, as not all sequences are acquired by default.

CONCLUSION

Our radiomics pilot to quantitatively assess T2-weighted MRI shows the potential to distinguish malignant from benign primary solid liver lesions in patients with noncirrhotic livers, both in internal validation and in two external validations based on heterogeneous multicenter data. However, the current performance is likely not sufficient yet, and further improvements are warranted, including extension to other phenotypes and combination with other MRI sequences. Pending further improvements, our model may serve as a robust, noninvasive, and low-cost aid for radiologists to diagnose liver cancer.

ETHICAL APPROVAL

The study protocol was reviewed and approved by the medical ethical committee of the Erasmus MC (Rotterdam, the Netherlands; MEC-2017-1035), Maastricht UMC+ (Maastricht, the Netherlands; METC 2018-0742), and Hôpital Beaujon (Paris, France; N° 2018-002). Informed consent was waived due to the use of retrospective, anonymized data. The study protocol conformed to the ethical guidelines of the 1975 Declaration of Helsinki.

AUTHOR CONTRIBUTIONS

M.P.A.S., R.L.M., V.V., M.R., W.J.N., S.K., and M.G.T. provided the conception and design of the study. M.P.A.S., R.L.M., Y.P., J.V., J.I., R.A.d.M., M.D., and M.G.T. acquired the data. M.P.A.S., S.K., and M.G.T. analyzed and interpreted the data. M.P.A.S. created the software. M.P.A.S., S.K., and M.G.T. drafted the article. All authors read and approved the final manuscript.

DATA AVAILABILITY STATEMENT

The imaging and clinical research data for part of this study (data set A) have been made publicly available: https://doi.org/10.1101/2021.08.19.21262238. Programming code is available on Zenodo at DOI https://doi.org/10.5281/zenodo.5175705.

DECLARATION OF COMPETING INTEREST

W.J.N. is the founder and shareholder of Quantib BV. The other authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

ACKNOWLEDGMENTS

M.P.A.S. acknowledges funding from the research program STRaTeGy (project number 14929-14930), which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). This study is supported by EuCanShare and EuCanImage (European Union's Horizon 2020 research and innovation program under grant agreement numbers 825903 and 952103, respectively). This work was partially carried out on the Dutch national e-infrastructure with the support of SURF Cooperative.

APPENDIX A. SUPPORTING INFORMATION

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.acra.2023.07.024.

REFERENCES

1. Sung H, Ferlay J, Siegel RL, et al. Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 2021; 71:209–249.
2. European Association for the Study of the Liver (EASL). EASL Clinical Practice Guidelines on the management of benign liver tumours. J Hepatol 2016; 65(2):386–398.
3. European Association for the Study of the Liver (EASL). EASL Clinical Practice Guidelines: management of hepatocellular carcinoma. J Hepatol 2018; 69:182–236.
4. Oka H, Kurioka N, Kim K, et al. Prospective study of early detection of hepatocellular carcinoma in patients with cirrhosis. Hepatology 1990; 12:680–687.
5. American College of Radiology Liver Reporting & Data System (LI-RADS). Available at: ⟨https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS⟩. Accessed September 8, 2022.
6. WHO Classification of Tumours Editorial Board. Digestive system tumours. 5th edn Lyon (France): International Agency for Research on Cancer; 2019.
7. Barth BK, Donati OF, Fischer MA, et al. Reliability, validity, and reader acceptance of LI-RADS-an in-depth analysis. Acad Radiol 2016; 23:1145–1153.
8. Desai A, Sandhu S, Lai JP, et al. Hepatocellular carcinoma in non-cirrhotic liver: a comprehensive review. World J Hepatol 2019; 11:1–18.
9. Song J, Yin Y, Wang H, et al. A review of original articles published in the emerging field of radiomics. Eur J Radiol 2020; 127:108991.
10. Saini A, Breen I, Pershad Y, et al. Radiogenomics and radiomics in liver cancers. Diagnostics 2019; 9:4.
11. Oestmann PM, Wang CJ, Savic LJ, et al. Deep learning–assisted differentiation of pathologically proven atypical and typical hepatocellular carcinoma (HCC) versus non-HCC on contrast-enhanced MRI of the liver. Eur Radiol 2021; 31:4981–4990.
12. Starmans MPA, Miclea RL, van der Voort SR, et al. Classification of malignant and benign liver tumors using a radiomics approach. SPIE Med Imaging 2018: Image Process 2018; 10574:105741D. https://doi.org/10.1117/12.2293609
13. Jansen MJA, Kuijf HJ, Veldhuis WB, et al. Automatic classification of focal liver lesions based on MRI and risk factors. PLoS One 2019; 14:e0217053.
14. Gatos I, Tsantis S, Karamesini M, et al. Focal liver lesions segmentation and classification in nonenhanced T2-weighted MRI. Med Phys 2017; 44:3695–3705.
15. Zhen S-h, Cheng M, Tao Y-b, et al. Deep learning for accurate diagnosis of liver tumor based on magnetic resonance imaging and clinical data. Front Oncol 2020; 10:680.
16. Hamm CA, Wang CJ, Savic LJ, et al. Deep learning for liver tumor diagnosis part I: development of a convolutional neural network classifier for multi-phasic MRI. Eur Radiol 2019; 29:3338–3347.
17. Kim Y, Furlan A, Borhani AA, et al. Computer-aided diagnosis program for classifying the risk of hepatocellular carcinoma on MR images following liver imaging reporting and data system (LI-RADS). J Magn Reson Imaging 2018; 47:710–722.
18. Wakabayashi T, Ouhmich F, Gonzalez-Cabrera C, et al. Radiomics in hepatocellular carcinoma: a quantitative review. Hepatol Int 2019; 13:546–559.
19. Ziol M, Pote N, Amaddeo G, et al. Macrotrabecular-massive hepatocellular carcinoma: a distinctive histological subtype with clinical relevance. Hepatology 2018; 68:103–112.
20. Canellas R, Burk KS, Parakh A, et al. Prediction of pancreatic neuroendocrine tumor grade based on CT features and texture analysis. AJR Am J Roentgenol 2018; 210:341–346.
21. Bise S, Frulio N, Hocquelet A, et al. New MRI features improve subtype classification of hepatocellular adenoma. Eur Radiol 2019; 29:2436–2447.
22. van Aalten SM, Thomeer MGJ, Terkivatan T, et al. Hepatocellular adenomas: correlation of MR imaging findings with pathologic subtype classification. Radiology 2011; 261:172–181.
23. Starmans MPA, Timbergen MJM, Vos M, et al. The WORC* database: MRI and CT scans, segmentations, and clinical labels for 930 patients from six radiomics studies. medRxiv 2021. https://doi.org/10.1101/2021.08.19.21262238:2021.2008.2019.21262238
24. Nagtegaal ID, Odze RD, Klimstra D, et al. The 2019 WHO classification of tumours of the digestive system. Histopathology 2020; 76:182–188.
25. Vilgrain V. Focal nodular hyperplasia. Eur J Radiol 2006; 58:236–245.
26. Befeler AS, Di Bisceglie AM. Hepatocellular carcinoma: diagnosis and treatment. Gastroenterology 2002; 122:1609–1619.
27. Bedogni G, Nobili V, Tiribelli C. Epidemiology of fatty liver: an update. World J Gastroenterol 2014; 20:9050–9054.
28. Zou KH, Warfield SK, Bharatha A, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 2004; 11:178–189.
29. Starmans MPA, van der Voort SR, Phil T, et al. Reproducible radiomics through automated machine learning validated on twelve clinical applications. arXiv 2021. https://doi.org/10.48550/arXiv.2108.08618:2108.08618
30. Starmans MPA, Van der Voort SR, Phil T, et al. Workflow for Optimal Radiomics Classification (WORC). Zenodo 2018. https://doi.org/10.5281/zenodo.3840534. Available at: ⟨https://github.com/MStarmans91/WORC⟩. Accessed September 8, 2022.
31. Starmans MPA. LiverRadiomics. Zenodo 2023. https://doi.org/10.5281/zenodo.5175705. Available at: ⟨https://github.com/MStarmans91/LiverRadiomics⟩. Accessed March 22, 2023.
32. Picard RR, Cook RD. Cross-validation of regression models. J Am Stat Assoc 1984; 79:575–583.
33. Nadeau C, Bengio Y. Inference for the generalization error. Mach Learn 2003; 52:239–281.
34. Efron B, Tibshirani R. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Stat Sci 1986; 1:54–75.
35. Macskassy SA, Provost F, Rosset S. ROC confidence bands: an empirical evaluation. Proceedings of the 22nd International Conference on Machine Learning 2005:537–544. https://doi.org/10.1145/1102351.1102419
36. McHugh ML. Interrater reliability: the kappa statistic. Biochem Med 2012; 22:276–282.
37. Schober P, Boer C, Schwarte LA. Correlation coefficients: appropriate use and interpretation. Anesth Analg 2018; 126:1763–1768.
38. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988; 44:837–845.
39. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017; 30.
40. Xiao X, Zhao J, Li S. Task relevance driven adversarial learning for simultaneous detection, size grading, and quantification of hepatocellular carcinoma via integrating multi-modality MRI. Med Image Anal 2022; 81:102554.
41. Bluemke DA, Moy L, Bredella MA, et al. Assessing radiology research on artificial intelligence: a brief guide for authors, reviewers, and readers—From the Radiology Editorial Board. Radiology 2019; 294:487–489.
42. Lambin P, Leijenaar RTH, Deist TM, et al. Radiomics: the bridge between medical imaging and personalized medicine. Nat Rev Clin Oncol 2017; 14:749–762.
43. Welch ML, McIntosh C, Haibe-Kains B, et al. Vulnerabilities of radiomic signature development: the need for safeguards. Radiother Oncol 2019; 130:2–9.
44. Gul S, Khan MS, Bibi A, et al. Deep learning techniques for liver and liver tumor segmentation: a review. Comput Biol Med 2022; 147:105620.
45. Vos M, Starmans MPA, Timbergen MJM, et al. Radiomics approach to distinguish between well differentiated liposarcomas and lipomas on MRI. Br J Surg 2019; 106:1800–1809.