Radiology

Repeatability of Diagnostic Features and Scoring Systems for Hepatocellular Carcinoma by Using MR Imaging[1]

Matthew S. Davenport, MD
Shokoufeh Khalatbari, MS
Peter S. C. Liu, MD
Katherine E. Maturen, MD
Ravi K. Kaza, MD
Ashish P. Wasnik, MD
Mahmoud M. Al-Hawary, MD
Daniel I. Glazer, MD
Erica B. Stein, MD
Jeet Patel, MD
Deepak K. Somashekar, MD
Benjamin L. Viglianti, MD, PhD
Hero K. Hussain, MD

Purpose: To determine for expert and novice radiologists repeatability of major diagnostic features and scoring systems (ie, Liver Imaging Reporting and Data System [LI-RADS], Organ Procurement and Transplantation Network [OPTN], and American Association for the Study of Liver Diseases [AASLD]) for hepatocellular carcinoma (HCC) by using magnetic resonance (MR) imaging.

Materials and Methods: Institutional review board approval was obtained and patient consent was waived for this HIPAA-compliant, retrospective study. The LI-RADS discussed in this article refers to version 2013.1. Ten blinded readers reviewed 100 liver MR imaging studies that demonstrated observations preliminarily assigned LI-RADS scores of LR1–LR5. Diameter and major HCC features (arterial hyperenhancement, washout appearance, pseudocapsule) were recorded for each observation. LI-RADS, OPTN, and AASLD scores were assigned. Interreader agreement was assessed by using intraclass correlation coefficients and κ statistics. Scoring rates were compared by using McNemar test.

Results: Overall interreader agreement was substantial for arterial hyperenhancement (0.67 [95% confidence interval {CI}: 0.65, 0.69]), moderate for washout appearance (0.48 [95%CI: 0.46, 0.50]), moderate for pseudocapsule (0.52 [95% CI: 050, 0.54]), fair for LI-RADS (0.35 [95% CI: 0.34, 0.37]), fair for AASLD (0.39 [95% CI: 0.37, 0.42]), and moderate for OPTN (0.53 [95% CI: 0.51, 0.56]). Agreement for measured diameter was almost perfect (range, 0.95–0.97). There was substantial agreement for most scores consistent with HCC. Experts agreed significantly more than did novices and were significantly more likely than were novices to assign a diagnosis of HCC ($P < .001$).

Conclusion: Two of three major features for HCC (washout appearance and pseudocapsule) have only moderate interreader agreement. Experts and novices who assigned scores consistent with HCC had substantial but not perfect agreement. Expert agreement is substantial for OPTN, but moderate for LI-RADS and AASLD. Novices were less consistent and less likely to diagnose HCC than were experts.

©RSNA, 2014

Online supplemental material is available for this article.

[1] From the Department of Radiology, University of Michigan Health System, 1500 E Medical Center Dr, Ann Arbor, MI 48109. Received August 20, 2013; revision requested October 7; revision received November 29; accepted December 12; final version accepted December 26. **Address correspondence to** M.S.D. (e-mail: *matdaven@med.umich.edu*).

An imaging-based diagnosis of hepatocellular carcinoma (HCC) is accepted as a surrogate for tissue confirmation in patients with chronic liver disease when classic criteria are met (arterial hyperenhancement, washout appearance, pseudocapsule appearance, and size ≥2 cm) (1–5). However, there is inconsistency in liver lesion reporting between radiologists, which can vary by experience level, local medical culture, differences in interpretation, and personal preference (2,5,6). Because tissue confirmation is often waived and major management decisions (eg, liver transplantation) are made when HCC is diagnosed with imaging, there is a need to provide reliable, consistent, and accurate reports (1,2,7–9). Patients can undergo unnecessary and morbid therapy if HCC is erroneously diagnosed, which can affect not only the index patient, but can have a profound indirect effect on unrelated patients because of a waste of scarce resources (eg, unnecessary liver transplantation) (5).

Recently, substantial efforts resulted in a series of structured reporting systems based on imaging to standardize the characterization and reporting of imaging observations made in the setting of chronic liver disease (Liver Imaging Reporting and Data System [LI-RADS] version 2013.1 [6], revised Organ Procurement and Transplantation Network [OPTN] criteria [1], American Association for the Study of Liver Diseases [AASLD] practice guideline [3]). The intention of these systems is twofold: to ensure 100% specificity when HCC is diagnosed with imaging, and to improve the coherency and consistency of radiology reports. If these two goals are to be met, the system that is used must have a high degree of interreader agreement. To date, this has not been well established.

The purpose of this study was to determine for expert and novice readers the repeatability of major diagnostic features and scoring systems (ie, LI-RADS, OPTN, and AASLD) for HCC by using magnetic resonance (MR) imaging.

Advances in Knowledge

- There was substantial to moderate repeatability of major diagnostic features for hepatocellular carcinoma (HCC) (arterial hyperenhancement, 0.67 [95% confidence interval {CI}: 0.65, 0.69]; washout appearance, 0.48 [95% CI: 0.46, 0.50]; pseudocapsule, 0.52 [95% CI: 050, 0.54]).
- Novices are less likely than experts to assign a diagnosis of HCC (Liver Imaging Reporting and Data System [LI-RADS] score of LR5, 17% [85 of 500] vs 28% [140 of 500], respectively; $P < .001$; Organ Procurement and Transplantation Network [OPTN] score of 5, 17% [83 of 500] vs 29% [143 of 500], respectively; $P < .001$).
- Overall interreader agreement was significantly better for OPTN (0.53 [95% CI: 0.51, 0.56]) than LI-RADS (0.35 [95% CI: 0.34, 0.37]) because of the weak agreement for intermediate-risk LI-RADS categories (eg, LR2, 0.11 [95% CI: 0.08, 0.14]; LR3, 0.26 [95% CI: 0.23, 0.29]; LR4, 0.28 [95% CI: 0.25, 0.31]).
- Interreader agreement for size (diameter) is almost perfect for all imaging phases after administration of contrast material and reader experience levels, both between readers (intraclass correlation coefficient range, 0.95–0.97) and within patients (intraclass correlation coefficient range, 0.94–0.98).
- The use of ancillary features to adjust LI-RADS scores is significantly associated with greater variability in interreader agreement ($P = .002$).

Implications for Patient Care

- Two of three major imaging-based diagnostic criteria for HCC, washout appearance and pseudocapsule, have only moderate interreader agreement, which represents a challenge for the noninvasive diagnosis of HCC.
- The LI-RADS system has fair-to-slight interreader agreement for intermediate-risk categories (LR2–LR4) and may need further refinement.

Materials and Methods

Institutional review board approval was obtained and patient consent waived for this Health Insurance Portability and Accountability Act–compliant, retrospective study. The LI-RADS discussed in this article refers to version 2013.1.

Subjects

A fellowship-trained abdominal radiologist (M.S.D., 2 years of experience) reviewed dynamic contrast agent–enhanced liver MR examinations performed on adult patients with chronic hepatitis or cirrhosis by using gadobenate dimeglumine (MultiHance; Bracco

Published online before print
10.1148/radiol.14131963 **Content code:** GI

Radiology 2014; 272:132–142

Abbreviations:
AASLD = American Association for the Study of Liver Diseases
CI = confidence interval
HCC = hepatocellular carcinoma
LI-RADS = Liver Imaging Reporting and Data System
OPTN = Organ Procurement and Transplantation Network

Author contributions:
Guarantor of integrity of entire study, M.S.D.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; literature research, M.S.D., H.K.H.; clinical studies, M.S.D., P.S.C.L., R.K.K., A.P.W., M.M.A., J.P., B.L.V., H.K.H.; experimental studies, D.I.G., B.L.V.; statistical analysis, M.S.D., S.K.; and manuscript editing, M.S.D., S.K., P.S.C.L., K.E.M., M.M.A., D.I.G., J.P., B.L.V., H.K.H.

Funding:
This research was supported by the National Institutes of Health (grant number UL1TR000433).

Conflicts of interest are listed at the end of this article.

See also the editorial by Baron in this issue.

Diagnostics, Princeton, NJ) between February 25, 2011, and January 12, 2013 (*n* = ~500). The radiologist identified and assigned a preliminary LI-RADS score to hepatic observations until there were 20 observations for each LI-RADS score from LR1 to LR5. The first eligible observations were selected until the quota had been filled. No patients who had undergone locoregional therapy with transcatheter arterial chemoembolization or hepatic radiation therapy before the study were included. Patients who had undergone hepatic thermal ablation before the study period were permitted for inclusion as long as the index observation was spatially distant from the thermal ablation cavity. We did not assess hepatobiliary phase imaging because it is not considered in the LI-RADS or OPTN classification schemes. Only one observation per MR examination was included, which resulted in 100 observations in 100 MR examinations in 95 patients. The preliminary LI-RADS scores were as follows: LR1, 20 preliminary scores; LR2, 20 preliminary scores; LR3, 20 preliminary scores; LR4, 20 preliminary scores; LR5, 20 preliminary scores. The study cohort consisted of 63 men (mean age, 61 years; age range, 35–82 years) and 32 women (mean age, 59 years; age range, 31–80 years). All patients in the study group had chronic viral hepatitis and/or cirrhosis.

Liver MR Protocol

All liver MR examinations included in this study were performed by using either 1.5-T or 3-T MR imagers (Achieva XR, Philips Healthcare, Best, the Netherlands; Ingenia, Philips Healthcare; LX Signa Excite 2, GE Healthcare, Milwaukee, Wis; HD, GE Healthcare), and used the following pulse sequences and parameters: breath-hold coronal single-shot T2-weighted turbo spin echo (repetition time msec/echo time msec, shortest/255; flip angle, 90°; section thickness, 8 mm; intersection gap, 0 mm; field of view, abdomen); breath-held axial single-shot turbo spin echo (shortest/235; flip angle, 90°; section thickness, 8 mm; intersection gap, 0 mm; field of view, liver); breath-hold axial T1-weighted dual-echo gradient-recalled echo (185/2.3, 4.6; flip angle, 70°; section thickness, 8 mm; intersection gap, 0 mm; field of view, liver); respiratory-triggered axial fat saturation T2-weighted fast spin echo (shortest/90; flip angle, 90°; section thickness, 8 mm; intersection gap, 0 mm; field of view, liver; spectral presaturation with inversion recovery fat saturation); respiratory-triggered axial diffusion-weighted imaging (echo-planar imaging; *b* values, 0 and 800 sec/mm^2; section thickness, 6 mm; intersection gap, 0 mm; field of view, liver); breath-hold axial T1-weighted fat-saturated three-dimensional precontrast and dynamic postcontrast (arterial, 20–30 seconds; venous, 60–90 seconds; extracellular, 120–150 seconds; delayed, 180–210 seconds) gradient echo (1.3/3.6; flip angle, 10°–12°; acquisition time, ~20 seconds; section thickness, 4 mm, interpolated to 2 mm; field of view, liver, spectral adiabatic inversion recovery fat saturation). Arterial phase timing was based on manual fluoroscopic (Acheiva XR, Philips Healthcare; Ingenia, Philips Healthcare) or automated contrast material bolus tracking (SmartPrep; GE Healthcare). We used the contrast agent gadobenate dimeglumine (MultiHance; Bracco Diagnostics, Princeton NJ), and the administered dose was selected according to patient weight (0.1 mmol/kg; maximum dose, 20 mL).

Image Processing

All patient identifiers were removed from each MR examination associated with a study observation by using the Radiological Society of North America digital randomization tool (10), and they were networked to a picture archiving and communications workstation (Horizon Medical Imaging PACS; McKesson, Richmond, Canada) under a dummy identifier. When one or more comparison studies were available (64 of 100 studies; 48 MR abdomen studies, 16 computed tomographic [CT] abdomen studies), the most relevant comparison used to assign the preliminary LI-RADS score was similarly deidentified and migrated to the research picture archiving and communications workstation under the same dummy identifier. All pulse sequences for both the index and comparison study (if any) were migrated and made available for review.

Observation Atlas

Single images that best depicted each observation (*n* = 100) were captured and stored in a digital atlas (PowerPoint; Microsoft Corporation, Redmond, Wash) to guide the reviewers and to allow for a targeted assessment of individual observations. The image that most clearly showed each observation was chosen, regardless of MR sequence type or postcontrast phase. Each slide of the atlas contained one representative image that depicted the targeted observation, an arrow that highlights the observation, the date of the study, the date and type of comparison study (if any), and the deidentified observation code to reference the study on the institutional research picture archiving and communications system. No patient identifiers were included in the atlas. The only link between the deidentified images for review and the deidentified atlas was the randomized study observation code.

Image Review

For this study, five fellowship-trained radiologists (P.S.C.L., K.E.M., R.K.K., A.P.W., M.M.A., 6–11 years of experience after fellowships) at a liver transplantation center and five novice radiology residents (D.I.G., E.B.S., J.P., D.K.S., B.L.V.) at the same center interpreted MR examinations. Before the beginning of the study, each reader (*n* = 10) was provided 1 hour of lecture-based and hands-on instruction that explained in detail each liver observation scoring system (LI-RADS, OPTN, AASLD) to be used in this study, with emphasis on the similarities and differences among them. The definitions of arterial hyperenhancement, washout appearance, and pseudocapsule appearance were provided by using the LI-RADS definitions (6). Differences in size measurements between the scoring systems were explained (eg, diameter measured on late arterial or early venous imaging for OPTN [1] vs diameter measured where observation is best indicated while avoiding the arterial phase if possible for LI-RADS [6]). The distinctions between OPTN 5A-g observations (1) and LR4A observations (6) were described. The readers were advised that ancillary criteria for HCC (defined in the LI-RADS glossary) may be used in the LI-RADS system (Fig E1 [online]), but they cannot be used for an observation from patient classification of LR4 or less to LR5 (6). In addition, each reader was provided 10 practice cases in a format identical to the study cases to be used for training purposes. Informal feedback was provided to the readers when questions arose during the training set. These steps were taken to minimize the effects of training bias during the study. The training cases were not included in the study group.

Each reader was provided access to a shared networked folder that contained a blank reader sheet, the study observation atlas, the official LI-RADS atlas and glossary published by the American College of Radiology (6), the current OPTN liver lesion classification system (1), and a recent publication of the AASLD schema (3). The images in the observation atlas were used as a roadmap to direct readers to the target observations. The complete MR examinations (not the single images from the observation atlas) were evaluated during the image review process. Table E1 (online) describes the AASLD schema assessed in our study. Table E2 (online) describes the OPTN guideline.

For each observation, each reader (who was blinded to other readers, the preliminary LI-RADS score, the distribution of preliminary LI-RADS scores, and patient history) measured the maximum diameter of the observation in each postcontrast phase (arterial, venous, extracellular, delayed), recorded the presence of major features for HCC (arterial hyperenhancement, washout appearance, pseudocapsule), and assigned scores according to each of the three systems (LI-RADS, OPTN, and AASLD). If the observation could not be measured on one or more phases, the reader was instructed not to record a size for those phases. AASLD does not have an official scoring system, so a scoring system was adapted from Bruix et al (3), where 1 = benign, 2 = follow-up, 3 = biopsy (>1-cm arterial-phase hyperenhanced observation without washout appearance), 4 = definite HCC (>1-cm arterial-phase hyperenhanced observation with washout appearance) (Table E1 [online]). For each LI-RADS score, the reader indicated whether LI-RADS ancillary features for HCC (6) were used to assign that score. Expert consensus for any given score or scoring system combination was considered if at least three of the five expert readers agreed on the same score.

Clinical Outcome

Whether follow-up, biopsy, resection, or treatment was necessary for the 100 observations that were interpreted by the readers was determined during retrospective chart review by a radiologist (M.S.D.). The median imaging follow-up period was 8 months (range, 0–84 months). Observations that had been considered consistent with HCC without tissue confirmation were presented to an institutional multidisciplinary liver tumor board before therapy.

Observations were arranged retrospectively into the following general categories based on the original prospective management: *(a)* not definite HCC according to imaging and clinical criteria; *(b)* presumed to be HCC according to imaging and clinical criteria without tissue confirmation (and with subcategorization based on the initial subsequent treatment: thermal ablation, chemoembolization, yttrium 90, sorafenib, stereotactic radiation therapy, hospice); *(c)* HCC confirmed by histologic analysis; *(d)* benign mass confirmed by histologic analysis; or *(e)* cholangiocarcinoma confirmed by histologic analysis. Retrospective reader scores were stratified overall and by experience level regarding the original prospective management.

Data Analysis

Continuous variables were summarized by using means and ranges. Categorical variables were summarized with counts and proportions. Intraclass correlation coefficients and 95% confidence intervals (CIs) were used to assess the interreader agreement as well as interphase agreement for liver observation size (maximum diameter). Mixed unconditional mean models were used for intraclass correlation coefficient calculations adjusted between readers and between subject correlations, as appropriate. Multirater Fleiss κ statistics and 95% CIs were used to assess the interreader agreement for scoring systems and major diagnostic features of HCC (ie, arterial hyperenhancement, washout appearance, and pseudocapsule). Bonferroni correction was applied to the κ statistic CIs of the three major scoring systems stratified by experience level. The CIs that were related to major diagnostic features of HCC were not adjusted because they were not involved in hypothesis testing. Other reported tests are considered exploratory.

κ results were stratified qualitatively by score (slight agreement, 0.01–0.20; fair agreement, 0.21–0.40; moderate agreement, 0.41–0.60; substantial agreement, 0.61–0.80; almost perfect agreement, 0.81–0.99 [11]). Scoring was compared by using McNemar test. To evaluate the effect of use of ancillary imaging features for HCC on LI-RADS scores, the variability of LI-RADS scores within 10 reader observations was assessed by analyzing the standard deviation of LI-RADS scoring as a dependent variable in a regression model. Two-tailed *P* values less than .05 indicated statistical significance. Data analysis was performed with statistical software (SAS 9.2; SAS Institute, Cary, NC).

Results

The distribution of LI-RADS, OPTN, and AASLD scores per reader is shown in Table 1. By expert consensus, none of the MR studies were nondiagnostic (ie, OPTN 0). Expert consensus was not reached for 19 LI-RADS observations, 19 OPTN

Table 1

Number of Observations in the Dataset Stratified by Expert Consensus, Individual Reader, and Scoring System Assignment

Score	Expert Consensus	Reader 1	Reader 2	Reader 3	Reader 4	Reader 5	Reader 6	Reader 7	Reader 8	Reader 9	Reader 10
LI-RADS scores											
LR1	16	28	17	14	15	20	8	18	14	17	15
LR2	1	15	10	9	5	13	11	17	26	13	23
LR3	26	15	27	31	31	18	29	35	26	34	18
LR4A	9	10	12	8	14	19	17	6	11	15	14
LR4B	2	7	7	6	7	2	17	8	9	4	10
LR5A	4	6	3	6	9	5	6	3	1	4	5
LR5B	20	16	18	25	13	18	12	10	12	12	15
LR5V	3	3	6	1	6	5	0	3	1	1	0
No consensus	19	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
OPTN scores											
0	0	0	1	0	0	0	0	3	0	0	0
Deferred to LI-RADS*	56	76	72	67	70	71	83	81	86	83	81
5A	4	5	2	7	9	5	5	3	0	3	3
5A-g	0	1	1	0	2	0	0	0	1	1	1
5B	14	12	17	17	12	15	9	9	6	11	10
5B-g	0	0	0	0	1	1	1	1	3	1	0
5X	7	6	7	9	6	8	2	3	4	1	5
No consensus	19	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
AASLD scores											
1	19	36	19	21	16	24	11	34	33	21	25
2	23	15	20	31	28	32	46	26	21	40	16
3	16	15	32	12	29	8	21	22	29	20	16
4	33	34	29	36	27	36	22	18	17	19	43
No consensus	9	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A

Note—Readers 1–5 were experts and readers 6–10 were novices. There were 100 observations in the dataset. An expert consensus score was assigned if at least three of five expert readers gave the same score or it was scored as no consensus.

* Scored as neither a category 5 nor category 0 OPTN observation; per OPTN guidelines, characterization is deferred to the LI-RADS system. AASLD scores are as follows: 1 = benign, 2 = follow, 3 = biopsy, 4 = HCC.

observations, and nine AASLD observations. Figure 1 and Figure 2 are examples of two study cases that caused substantial reader disagreement. Scores of LR2 were variably assigned (median number of LI-RADS scores of LR2 by reader per 100 observations, 13 [all readers], 10 [expert readers], 17 [novice readers]), but inconsistently so (ie, only one observation was considered LR2 by expert consensus). OPTN scores of 5A-g and 5B-g were rarely assigned (*n* = 0 with expert consensus; OPTN 5A-g range by reader, 0–2; OPTN 5B-g range by reader, 0–3). Novices were less likely than were experts to assign a diagnosis of HCC (LR5: 17% [85 of 500] vs 28% [140 of 500], respectively; $P < .001$; OPTN 5: 17% [83 of 500] vs 29% [143 of 500], respectively; $P < .001$; AASLD 4: 24% [119 of 500] vs 32% [161 of 500], respectively; $P < .001$). Novices were also less likely than were experts to assign a score that would contraindicate liver transplantation (LR5V: 1% [five of 500] vs 4% [21 of 500], respectively; $P < .001$; OPTN 5X: 3% [15 of 500] vs 7% [36 of 500], respectively; $P < .001$).

The details of the repeatability of the scoring systems are in Table 2. Experts were in agreement more than were novices for all scoring systems (LI-RADS: 0.43 [95% CI: 0.41, 0.45] vs 0.35 [95% CI: 0.33, 0.37], respectively; OPTN: 0.64 [95% CI: 0.60, 0.68] vs 0.50 [95% CI: 0.46, 0.54], respectively; AASLD: 0.46 [95% CI: 0.42, 0.50] vs 0.36 [95% CI: 0.32, 0.40], respectively). The OPTN system was significantly more repeatable (0.53 [95% CI: 0.51, 0.56]) than either LI-RADS (0.35 [95% CI: 0.34, 0.37]) or AASLD (0.39 [95% CI: 0.37, 0.42]). Interreader agreement for LI-RADS scores of LR2 (0.11 [95% CI: 0.08, 0.14]), LR3 (0.26 [95% CI: 0.23, 0.29]), and LR4 (0.28 [95% CI: 0.25, 0.31]) were significantly worse than the interreader agreement for LI-RADS scores of LR1 (0.54 [95% CI: 0.51, 0.57]) and LR5 (0.62 [95% CI: 0.59, 0.65]). Scores that were considered to be diagnostic of HCC had substantial repeatability between experts for all scoring systems (LI-RADS score of LR5: 0.71 [95% CI: 0.65, 0.77], OPTN 5: 0.70 [95% CI: 0.63, 0.76], AASLD 4: 0.61 [95% CI: 0.55, 0.67]). Table 3 shows the interreader agreement for all readers and scoring systems in the study group.

Figure 1

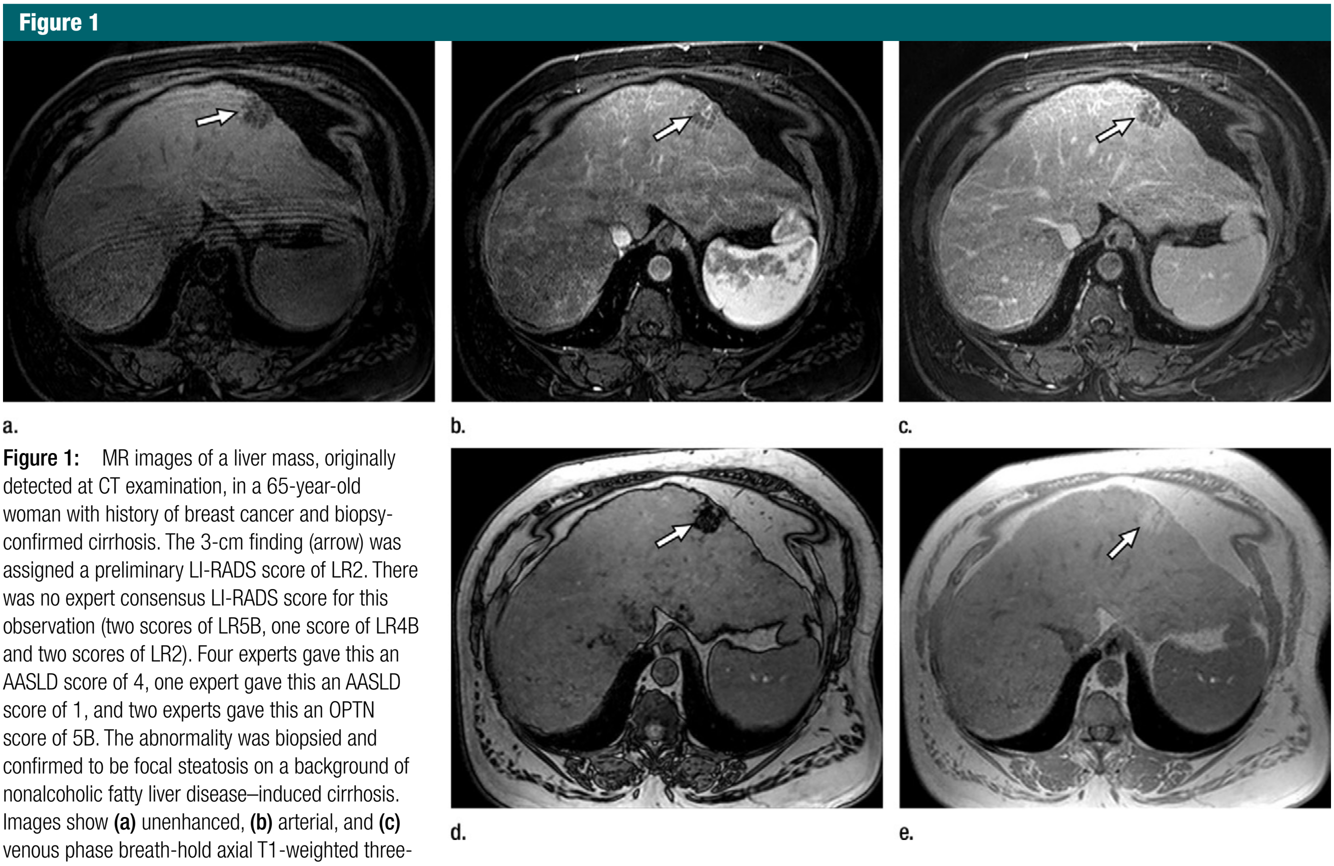

Figure 1: MR images of a liver mass, originally detected at CT examination, in a 65-year-old woman with history of breast cancer and biopsy-confirmed cirrhosis. The 3-cm finding (arrow) was assigned a preliminary LI-RADS score of LR2. There was no expert consensus LI-RADS score for this observation (two scores of LR5B, one score of LR4B and two scores of LR2). Four experts gave this an AASLD score of 4, one expert gave this an AASLD score of 1, and two experts gave this an OPTN score of 5B. The abnormality was biopsied and confirmed to be focal steatosis on a background of nonalcoholic fatty liver disease–induced cirrhosis. Images show **(a)** unenhanced, **(b)** arterial, and **(c)** venous phase breath-hold axial T1-weighted three-dimensional spoiled gradient-recalled-echo imaging with fat saturation, and **(d)** opposed-phase and **(e)** in-phase breath-hold axial T1-weighted two-dimensional dual-echo gradient-recalled-echo imaging with echo times of 2.3 msec and 4.6 msec, respectively.

Overall interreader agreement was substantial for arterial hyperenhancement (0.67 [95% CI: 0.65, 0.69]), moderate for washout appearance (0.48 [95% CI: 0.46, 0.50]), and moderate for pseudocapsule (0.52 [95% CI: 050, 0.54]) (Table 4). Expert-assigned size measurements were marginally but significantly more repeatable than novice-assigned size measurements for all postcontrast imaging phases, observation scores, and scoring systems (range of intraclass correlation coefficients for all readers and phases: 0.90–0.98). This persisted when only expert consensus LI-RADS scores of LR5A, LR5B, and LR5V were considered (range for novice readers, 0.93–0.96; range for expert readers, 0.97–0.98; range for all readers, 0.94–0.96). However, repeatability of size determination was almost perfect for all readers, regardless of experience level.

Measured diameters are stratified by expert consensus score in Table E3 [online]. There was almost perfect agreement for observation size within readers at both experience levels for observations across imaging phases after administration of contrast material (intraclass correlation coefficient range, 0.94–0.98) (Table E4 [online]), which indicated that the size recorded for an observation in one phase was fairly consistent across all imaging phases after administration of contrast material in which that observation was visible.

Ancillary imaging features for HCC were commonly used by all readers when assigning LI-RADS scores (Table E5 [online]), but experts used them significantly more often than did novices (25% [126 of 500] vs 18% [92 of 500]; *P* = .004). Overall use of ancillary features for LI-RADS scoring was associated with a significant increase in the variability of the assigned score (*P* = .002). For every additional reader that used ancillary features for any given observation, the standard deviation of LI-RADS scoring for that observation increased by 0.09.

The prospective clinical outcome is shown in Table E6 (online) and stratified by scoring system assignment. Of the liver observations that were determined by a combination of imaging and clinical criteria not to be consistent with HCC in actual clinical practice, 6% (16 of 280) of expert scores and 3% (seven of 280) of novice scores for those observations were LR5 and 6% (17 of 280) of expert scores and 3% (seven of 280) of novice scores were OPTN 5. Of the liver observations that were treated in actual clinical practice as consistent with HCC on the basis of imaging and clinical parameters, 0% (zero of 150) of expert

Figure 2

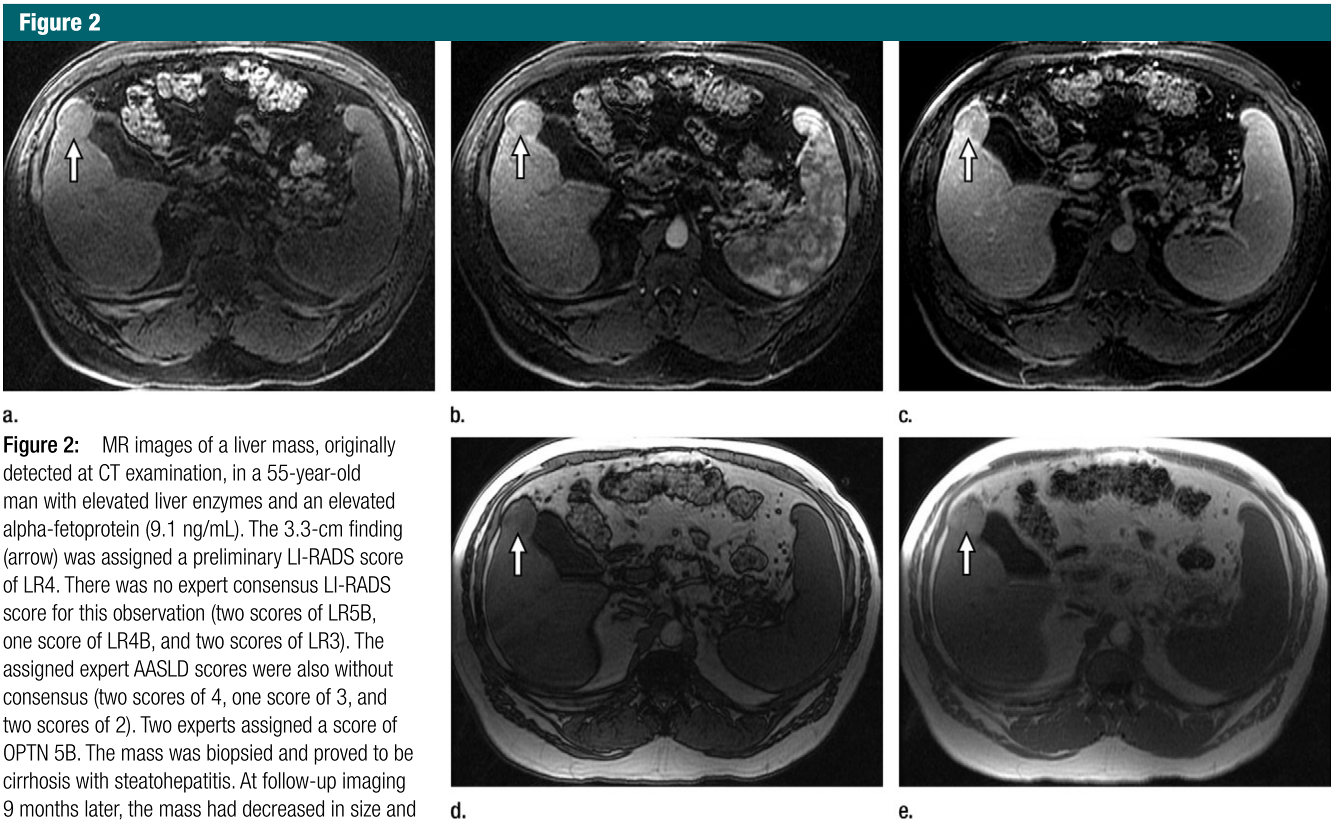

Figure 2: MR images of a liver mass, originally detected at CT examination, in a 55-year-old man with elevated liver enzymes and an elevated alpha-fetoprotein (9.1 ng/mL). The 3.3-cm finding (arrow) was assigned a preliminary LI-RADS score of LR4. There was no expert consensus LI-RADS score for this observation (two scores of LR5B, one score of LR4B, and two scores of LR3). The assigned expert AASLD scores were also without consensus (two scores of 4, one score of 3, and two scores of 2). Two experts assigned a score of OPTN 5B. The mass was biopsied and proved to be cirrhosis with steatohepatitis. At follow-up imaging 9 months later, the mass had decreased in size and the intracellular lipid within the mass had resolved. Images show **(a)** unenhanced, **(b)** arterial, and **(c)** venous phase breath-hold axial T1-weighted three-dimensional spoiled gradient-recalled-echo imaging with fat-saturation, and **(d)** opposed-phase and **(e)** in-phase breath-hold axial T1-weighted two-dimensional dual-echo gradient-recalled-echo imaging with echo times of 2.3 msec and 4.6 msec, respectively.

scores and 5% (seven of 150) of novice scores were either LR1 or LR2. No expert or novice assigned a low-risk LI-RADS score of LR1 or LR2 to a histologic-analysis confirmed HCC (zero of 40 expert scores; zero of 40 novice scores). Experts often assigned a high-risk LI-RADS score of LR4 (24% [six of 25]) or LR5 (16% [four of 25]) to observations that were shown to be benign through histologic analysis, while novice radiologists did so less often (LI-RADS score of LR4, 16% [four of 25]; LI-RADS score of LR5, 0% [zero of 25]). In general, experts assigned high-risk scores (LI-RADS scores of LR4 and LR5, OPTN score of 5, and AASLD score of 4) more often than did novices (Table E6 [online]), and were less likely than were novices to assign a low-risk score (LI-RADS scores of LR1 or LR2) to observations clinically treated as HCC.

Discussion

Our data demonstrate that there is substantial variation in liver observation reporting by both experts and novices when standardized reporting schema are used. Although experts consistently had significantly higher agreement than novices, only experts who used OPTN had substantial interreader agreement. The remaining scoring system and experience level combinations had agreement that was only fair to moderate. This appeared to be directly related to the following key issues: *(a)* imperfect interreader repeatability of qualitative major diagnostic features for HCC (ie, arterial hyperenhancement [substantial repeatability], washout appearance [moderate repeatability], pseudocapsule [moderate repeatability]); *(b)* fair-to-slight interreader repeatability of intermediate-suspicion categories in the LI-RADS system (LR2, LR3, LR4) and AASLD system (AASLD scores of 2 and 3); and *(c)* variability in interreader repeatability (*P* = .002) when subjective ancillary features for HCC were used.

The OPTN system is designed to indicate whether a patient is suited for transplantation on the basis of an imaging diagnosis of HCC, while the AASLD and LI-RADS systems add additional frameworks for the management of findings that are not consistent with HCC (1,3,6). It is critical that OPTN scores of 5 always indicate HCC because transplantation decisions are made based on those designations. Our data show that there was substantial but not perfect repeatability for most scores considered diagnostic of HCC across the three scoring systems, with smaller masses having less repeatability than larger masses.

Table 2

Repeatability of Liver Observation Scoring Systems Stratified by Experience Level

Characteristic	Novice Readers	Expert Readers	All Readers
HCC scoring systems			
LI-RADS score (overall)	0.35 (0.31, 0.38)	0.43 (0.41, 0.46)	0.35 (0.34, 0.37)
OPTN score	0.50 (0.45, 0.55)	0.64 (0.59, 0.69)	0.53 (0.51, 0.56)
AASLD score	0.36 (0.31, 0.41)	0.46 (0.41, 0.51)	0.39 (0.37, 0.42)
Observation, not diagnostic			
OPTN 0	N/A	N/A	N/A
Low-risk observation			
LR1	0.54 (0.48, 0.60)	0.65 (0.59, 0.71)	0.54 (0.51, 0.57)
LR2	0.11 (0.04, 0.17)	0.11 (0.05, 0.17)	0.11 (0.08, 0.14)
AASLD 1	0.40 (0.34, 0.47)	0.57 (0.51, 0.64)	0.47 (0.44, 0.50)
AASLD 2	0.29 (0.23, 0.35)	0.33 (0.27, 0.39)	0.31 (0.28, 0.34)
Intermediate-risk observation			
LR3	0.26 (0.20, 0.32)	0.31 (0.24, 0.37)	0.26 (0.23, 0.29)
AASLD 3	0.27 (0.21, 0.33)	0.26 (0.20, 0.32)	0.24 (0.21, 0.27)
High-risk observation			
LR4 (includes LR4A and LR4B)	0.34 (0.28, 0.40)	0.33 (0.27, 0.40)	0.28 (0.25, 0.31)
LR4A	0.39 (0.33, 0.45)	0.36 (0.30, 0.43)	0.34 (0.32, 0.38)
LR4B	0.34 (0.28, 0.41)	0.18 (0.11, 0.24)	0.23 (0.20, 0.26)
Observation indicates HCC			
LR5 (includes LR5A, LR5B, or LR5V)	0.62 (0.56, 0.68)	0.71 (0.65, 0.77)	0.62 (0.59, 0.65)
LR5A	0.37 (0.31, 0.43)	0.36 (0.30, 0.42)	0.36 (0.33, 0.39)
LR5B	0.60 (0.54, 0.66)	0.67 (0.61, 0.73)	0.59 (0.56, 0.62)
LR5V	0.19 (0.13, 0.25)	0.58 (0.52, 0.64)	0.40 (0.37, 0.43)
OPTN 5 (includes 5A, 5A-g, 5B, 5B-g, or 5X)	0.62 (0.56, 0.68)	0.70 (0.63, 0.76)	0.61 (0.58, 0.64)
OPTN 5A	0.27 (0.20, 0.33)	0.38 (0.31, 0.44)	0.32 (0.29, 0.34)
OPTN 5A-g	N/A	N/A	N/A
OPTN 5B	0.55 (0.49, 0.61)	0.69 (0.63, 0.75)	0.58 (0.55, 0.61)
OPTN 5B-g	N/A	N/A	N/A
OPTN 5X	0.42 (0.35, 0.48)	0.78 (0.71, 0.84)	0.57 (0.54, 0.60)
AASLD 4	0.48 (0.41, 0.54)	0.61 (0.55, 0.67)	0.52 (0.49, 0.55)

Note.—Data are κ statistics. Data in parentheses are 95% CIs. There were 100 observations, five expert readers, and five novice readers. N/A = no repeatability measurement was calculated because there were zero expert consensus observations with this designation.

The OPTN system was significantly more repeatable overall than both the LI-RADS and AASLD systems. This is because OPTN scores are basically restricted to those that indicate that an HCC is present (a designation shown in our study to have substantial repeatability). When only observations scored LR5 and AASLD 4 (ie, observations believed to be diagnostic of HCC) are considered, the repeatability is similar to that of OPTN 5 observations. The similarity between LI-RADS score of LR5 and OPTN score of 5 is not surprising because the category definitions are nearly identical. AASLD score of 4 had somewhat lower agreement, which is likely because it allows smaller (11–19 mm) arterial-phase hyperenhanced observations with washout appearance to be consistent with HCC without requiring a pseudocapsule or threshold growth (at least one of which is required in addition to the other features to meet criteria for either OPTN score of 5 or LI-RADS score of LR5).

Size was not a source of discrepancy between readers, and there was almost perfect agreement across all observation types for both reader experience levels. Interestingly, measurement sizes also were fairly similar within readers across imaging phases after administration of contrast material, which indicated that the size measured in one phase was fairly consistent with the size measured in other phases as long as the observation remained visible. This information has relevance for diameter measurements of liver observations, which is a key determinant of hepatic transplant eligibility with respect to Milan criteria (1,8,9). OPTN protocol requires that the measurement of a liver observation be made on the late hepatic arterial or early portal venous phase (1), while LI-RADS allows for the measurement to be made on the sequence where the observation is best outlined (6). Additionally, if the observation is best outlined on a sequence other than the late hepatic arterial phase, LI-RADS guidelines state that the late hepatic arterial phase should not be used for measurement purposes (6). The recommendations for diameter measurement between the two systems are nearly opposite, but this difference does not appear to play a substantial role in discrepant reporting. The method of size measurement is not explicitly specified for the AASLD system (3).

Although our study was not designed to directly assess the diagnostic performance of the various liver observation reporting systems, we did find some interesting results when we cross-referenced our retrospective findings with the original prospective management. Experts were significantly more likely to assign a retrospective HCC diagnosis (OPTN 5, AASLD 4, and LI-RADS scores of LR4 and LR5) than were novices, but this was true even for observations that were prospectively managed as benign. Experts were also significantly more likely than were novices to assign a score that would contraindicate liver transplantation. In general, experts behaved with greater sensitivity for HCC than novices. Because our study was not designed to primarily investigate clinical outcome, it is not clear whether these retrospective high-risk designations were more accurate than the prospective interpretations

Table 3

Interreader Agreement for Liver Observations Considered to Be Definite HCC by Applying LI-RADS, OPTN, or AASLD

Scoring System	Reader	Reader 2	Reader 3	Reader 4	Reader 5	Reader 6	Reader 7	Reader 8	Reader 9	Reader 10
LI-RADS 5	Reader 1	0.69 (0.53, 0.85)	0.73 (0.59, 0.88)	0.67 (0.50, 0.83)	0.77 (0.63, 0.91)	0.56 (0.37, 0.75)	0.61 (0.42, 0.79)	0.59 (0.40, 0.79)	0.70 (0.53, 0.87)	0.69 (0.51, 0.86)
OPTN 5	Reader 1	0.71 (0.55, 0.87)	0.73 (0.59, 0.88)	0.65 (0.48, 0.81)	0.72 (0.56, 0.87)	0.54 (0.34, 0.74)	0.63 (0.44, 0.82)	0.62 (0.43, 0.81)	0.73 (0.56, 0.89)	0.68 (0.50, 0.85)
AASLD 4	Reader 1	0.61 (0.44, 0.77)	0.82 (0.71, 0.94)	0.51 (0.33, 0.69)	0.65 (0.49, 0.81)	0.46 (0.28, 0.65)	0.55 (0.38, 0.72)	0.42 (0.23, 0.60)	0.58 (0.41, 0.75)	0.56 (0.40, 0.72)
LI-RADS 5	Reader 2	…	0.69 (0.53, 0.84)	0.62 (0.45, 0.80)	0.77 (0.63, 0.91)	0.52 (0.32, 0.71)	0.56 (0.37, 0.75)	0.55 (0.36, 0.74)	0.71 (0.55, 0.87)	0.48 (0.28, 0.67)
OPTN 5	Reader 2	…	0.72 (0.57, 0.86)	0.58 (0.41, 0.76)	0.75 (0.61, 0.90)	0.48 (0.28, 0.68)	0.56 (0.37, 0.75)	0.55 (0.36, 0.74)	0.71 (0.55, 0.87)	0.50 (0.30, 0.69)
AASLD 4	Reader 2	…	0.52 (0.35, 0.70)	0.70 (0.55, 0.86)	0.52 (0.35, 0.70)	0.45 (0.25, 0.65)	0.54 (0.35, 0.72)	0.45 (0.25, 0.64)	0.51 (0.32, 0.70)	0.36 (0.18, 0.54)
LI-RADS 5	Reader 3	…	…	0.67 (0.51, 0.83)	0.81 (0.68, 0.94)	0.43 (0.24, 0.62)	0.58 (0.40, 0.75)	0.51 (0.34, 0.69)	0.61 (0.44, 0.78)	0.54 (0.36, 0.72)
OPTN 5	Reader 3	…	…	0.65 (0.49, 0.81)	0.81 (0.69, 0.94)	0.43 (0.25, 0.62)	0.56 (0.39, 0.73)	0.50 (0.32, 0.67)	0.59 (0.42, 0.76)	0.54 (0.37, 0.72)
AASLD 4	Reader 3	…	…	0.56 (0.39, 0.73)	0.70 (0.55, 0.84)	0.48 (0.30, 0.66)	0.51 (0.34, 0.68)	0.44 (0.26, 0.61)	0.59 (0.43, 0.75)	0.52 (0.35, 0.69)
LI-RADS 5	Reader 4	…	…	…	0.70 (0.55, 0.86)	0.55 (0.37, 0.74)	0.66 (0.49, 0.83)	0.47 (0.28, 0.67)	0.63 (0.46, 0.81)	0.51 (0.32, 0.70)
OPTN 5	Reader 4	…	…	…	0.64 (0.47, 0.81)	0.48 (0.29, 0.67)	0.62 (0.44, 0.79)	0.44 (0.25, 0.63)	0.59 (0.42, 0.77)	0.49 (0.30, 0.68)
AASLD 4	Reader 4	…	…	…	0.47 (0.29, 0.65)	0.54 (0.35, 0.73)	0.69 (0.52, 0.86)	0.54 (0.35, 0.73)	0.50 (0.30, 0.69)	0.44 (0.27, 0.62)
LI-RADS 5	Reader 5	…	…	…	…	0.55 (0.37, 0.74)	0.60 (0.42, 0.78)	0.59 (0.41, 0.77)	0.69 (0.53, 0.85)	0.57 (0.38, 0.75)
OPTN 5	Reader 5	…	…	…	…	0.50 (0.31, 0.69)	0.58 (0.40, 0.76)	0.57 (0.39, 0.75)	0.67 (0.50, 0.83)	0.57 (0.38, 0.75)
AASLD 4	Reader 5	…	…	…	…	0.53 (0.35, 0.70)	0.46 (0.29, 0.64)	0.44 (0.26, 0.61)	0.54 (0.37, 0.71)	0.52 (0.35, 0.69)
LI-RADS 5	Reader 6	…	…	…	…	…	0.65 (0.44, 0.85)	0.63 (0.42, 0.84)	0.55 (0.33, 0.77)	0.55 (0.34, 0.76)
OPTN 5	Reader 6	…	…	…	…	…	0.60 (0.39, 0.81)	0.58 (0.36, 0.80)	0.50 (0.28, 0.73)	0.59 (0.39, 0.80)
AASLD 4	Reader 6	…	…	…	…	…	0.50 (0.29, 0.71)	0.59 (0.39, 0.79)	0.48 (0.27, 0.69)	0.46 (0.29, 0.62)
LI-RADS 5	Reader 7	…	…	…	…	…	…	0.61 (0.39, 0.83)	0.82 (0.66, 0.97)	0.66 (0.47, 0.85)
OPTN 5	Reader 7	…	…	…	…	…	…	0.61 (0.39, 0.83)	0.82 (0.66, 0.97)	0.69 (0.50, 0.88)
AASLD 4	Reader 7	…	…	…	…	…	…	0.55 (0.33, 0.77)	0.64 (0.44, 0.83)	0.41 (0.25, 0.57)
LI-RADS 5	Reader 8	…	…	…	…	…	…	…	0.66 (0.45, 0.86)	0.51 (0.29, 0.73)
OPTN 5	Reader 8	…	…	…	…	…	…	…	0.66 (0.45, 0.86)	0.53 (0.31, 0.75)
AASLD 4	Reader 8	…	…	…	…	…	…	…	0.59 (0.39, 0.80)	0.38 (0.22, 0.54)
LI-RADS 5	Reader 9	…	…	…	…	…	…	…	…	0.57 (0.36, 0.78)
OPTN 5	Reader 9	…	…	…	…	…	…	…	…	0.59 (0.39, 0.80)
AASLD 4	Reader 9	…	…	…	…	…	…	…	…	0.39 (0.22, 0.55)

Note.—Data are κ statistics and data in parentheses are 95% CI. Definite HCC was represented by LI-RADS scores of LR5A, LR5B, and LR5V; OPTN scores of 5A, 5A-g, 5B, 5B-g, and 5X; and AASLD score of 4.

Table 4

Repeatability of Quantitative and Qualitative Liver Observation Assessment Stratified by Experience Level

Characteristic	Novice Readers	Expert Readers	All Readers
Major HCC features (κ)			
Arterial hyperenhancement	0.67 (0.61, 0.73)	0.72 (0.66, 0.78)	0.67 (0.65, 0.69)
Washout appearance	0.56 (0.50, 0.62)	0.54 (0.48, 0.60)	0.48 (0.46, 0.50)
Pseudocapsule	0.52 (0.46, 0.58)	0.63 (0.57, 0.69)	0.52 (0.50, 0.54)
Intraclass correlation coefficients of size for all observations			
Arterial	0.96 (0.96, 0.97)	0.98 (0.98, 0.98)	0.97 (0.96, 0.98)
Venous	0.95 (0.94, 0.95)	0.98 (0.98, 0.98)	0.96 (0.95, 0.97)
Extracellular	0.94 (0.93, 0.95)	0.97 (0.97, 0.98)	0.95 (0.94, 0.96)
Delayed	0.94 (0.93, 0.94)	0.98 (0.97, 0.98)	0.95 (0.95, 0.96)
Largest size	0.96 (0.96, 0.97)	0.98 (0.98, 0.98)	0.97 (0.96, 0.97)
Intraclass correlation coefficients of size for LI-RADS scores of LR5A, LR5B, or LR5V			
Arterial	0.94 (0.93, 0.95)	0.97 (0.97, 0.98)	0.95 (0.95, 0.96)
Venous	0.95 (0.94, 0.95)	0.98 (0.97, 0.98)	0.96 (0.95, 0.97)
Extracellular	0.94 (0.94, 0.95)	0.97 (0.96, 0.97)	0.95 (0.94, 0.96)
Delayed	0.90 (0.89, 0.92)	0.97 (0.97, 0.98)	0.94 (0.92, 0.95)
Largest size	0.97 (0.97, 0.98)	0.97 (0.97, 0.98)	0.96 (0.96, 0.97)
Intraclass correlation coefficients of size for LI-RADS scores of LR4A, LR4B, LR5A, LR5B, or LR5V			
Arterial	0.92 (0.90, 0.93)	0.97 (0.97, 0.98)	0.94 (0.93, 0.96)
Venous	0.93 (0.92, 0.93)	0.98 (0.97, 0.98)	0.95 (0.93, 0.96)
Extracellular	0.92 (0.92, 0.93)	0.97 (0.96, 0.97)	0.94 (0.92, 0.96)
Delayed	0.90 (0.89, 0.92)	0.97 (0.97, 0.98)	0.94 (0.92, 0.96)
Largest size	0.96 (0.96, 0.97)	0.97 (0.97, 0.98)	0.96 (0.94, 0.97)
Intraclass correlation coefficients of size for OPTN scores of 5A, 5A-g, 5B, 5B-g, or 5X			
Arterial	0.94 (0.93, 0.95)	0.97 (0.97, 0.98)	0.95 (0.95, 0.96)
Venous	0.95 (0.94, 0.95)	0.98 (0.97, 0.98)	0.96 (0.95, 0.97)
Extracellular	0.94 (0.94, 0.95)	0.97 (0.96, 0.97)	0.95 (0.94, 0.96)
Delayed	0.90 (0.89, 0.92)	0.97 (0.97, 0.98)	0.93 (0.92, 0.95)
Largest size	0.97 (0.97, 0.98)	0.97 (0.97, 0.97)	0.96 (0.96, 0.97)
Intraclass correlation coefficients of size for AASLD scores of 3 or 4			
Arterial	0.95 (0.94, 0.96)	0.98 (0.97, 0.98)	0.96 (0.95, 0.97)
Venous	0.94 (0.94, 0.95)	0.98 (0.97, 0.98)	0.95 (0.94, 0.96)
Extracellular	0.94 (0.94, 0.95)	0.97 (0.97, 0.97)	0.94 (0.93, 0.95)
Delayed	0.94 (0.93, 0.95)	0.97 (0.97, 0.98)	0.94 (0.93, 0.95)
Largest size	0.97 (0.96, 0.97)	0.97 (0.97, 0.98)	0.96 (0.96, 0.97)

Note.—Data in parentheses are 95% CIs. There were five expert readers and five novice readers.

(ie, HCC was present but not recognized prospectively), or less so (ie, HCC was over-diagnosed in the retrospective review).

LI-RADS scores of LR2 were uncommonly assigned by experts and rarely agreed upon. The LR2 designation was not applied by experts to observations that were managed prospectively as HCC. Expert-assigned scores of LR3 were more common, and were assigned in cases where the observation was prospectively treated as benign (33% [102 of 305]) and where the observation was prospectively treated as HCC (11% [20 of 190]). Based on these data, retrospective expert-assigned LI-RADS scores of LR1 and LR2 indicated a negligible risk for HCC, while expert-assigned scores of LR3 and higher indicated a nontrivial risk of HCC.

Our study had several limitations. This study was designed to assess the repeatability of major diagnostic features and scoring systems for HCC. Therefore, follow-up data and histologic confirmation of the imaging findings are limited. A study that is dedicated to this goal, and has lengthier follow-up and greater histologic confirmation of findings is needed to permit stratification of HCC risk by LI-RADS designation (similar to the Breast Imaging Reporting and Data System [12]). The specific explanation or explanations for why individual readers disagreed on fundamental diagnostic criteria for HCC despite the use of clearly stated definitions (ie, LI-RADS glossary) is difficult to explain. However, this variation of subjective criteria interpretation is precisely the reason why standardized scoring systems for liver imaging have evolved. We show in our data that the use of similarly subjective minor ancillary criteria for HCC is inherently associated with greater variability in scoring repeatability. A study regarding the decision-making process when a radiologist determines whether a subjective feature is present would be enlightening. Some of the subcategories within the LI-RADS and OPTN scoring systems were not often scored by either expert or novice readers (eg, OPTN 5A-g, OPTN 5B-g, and LI-RADS scores of LR5A and LR5V). Therefore, these scores may be underrepresented in our dataset. Although the repeatability between readers in our study was substantial for both LR5 and OPTN 5, we may not be able to generalize the results to these less well-represented subcategories. Finally, our study only included MR examinations, and is not directly applicable to the assessment of liver observations with CT imaging.

In conclusion, two of three major qualitative MR imaging features for

HCC (washout appearance and pseudocapsule) have only moderate interreader agreement. Experts and novices who assigned scores that were consistent with HCC had substantial but not perfect agreement. Expert agreement was substantial for OPTN, but moderate for both LI-RADS and AASLD. This difference likely relates to the significantly poorer repeatability for scores that indicated an intermediate risk for HCC (ie, not definitively malignant and not definitively benign). Further refinement to the LI-RADS system, particularly for scores LR2, LR3, and LR4, is needed. The benefits of the use of subjective ancillary criteria for LI-RADS score assignment must be balanced against the cost of greater interreader variability.

Acknowledgments: We thank James Good, Senior Business Systems Analyst, and William Weadock, MD, for their help with image management and image anonymization.

Disclosures of Conflicts of Interest: M.S.D. No relevant conflicts of interest to disclose. **S.K.** No relevant conflicts of interest to disclose. **P.S.C.L.** No relevant conflicts of interest to disclose. **K.E.M.** No relevant conflicts of interest to disclose. **R.K.K.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: author received payment for lectures from GE Healthcare. Other relationships: none to disclose. **A.P.W.** No relevant conflicts of interest to disclose. **M.M.A.** No relevant conflicts of interest to disclose. **D.I.G.** No relevant conflicts of interest to disclose. **E.B.S.** No relevant conflicts of interest to disclose. **J.P.** No relevant conflicts of interest to disclose. **D.K.S.** No relevant conflicts of interest to disclose. **B.L.V.** No relevant conflicts of interest to disclose. **H.K.H.** Financial activities related to the present article: none to disclose. Financial activities not related to the present article: author is a consultant for Bayer Healthcare and Eovist Contrast; author receives payment for lectures from Eovist Contrast; author receives book royalties. Other relationships: none to disclose.

References

1. HRSA/OPTN. Policy 3.6 organ distribution: allocation of livers. http://optn.transplant.hrsa.gov/policiesAndBylaws/policies.asp. Published 2012. Accessed July 27, 2013.
2. Wald C, Russo MW, Heimbach JK, Hussain HK, Pomfret EA, Bruix J. New OPTN/UNOS policy for liver transplant allocation: standardization of liver imaging, diagnosis, classification, and reporting of hepatocellular carcinoma. Radiology 2013;266(2):376–382.
3. Bruix J, Sherman M; American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma: an update. Hepatology 2011;53(3):1020–1022.
4. Willatt JM, Hussain HK, Adusumilli S, Marrero JA. MR imaging of hepatocellular carcinoma in the cirrhotic liver: challenges and controversies. Radiology 2008;247(2):311–330.
5. Pomfret EA, Washburn K, Wald C, et al. Report of a national conference on liver allocation in patients with hepatocellular carcinoma in the United States. Liver Transpl 2010;16(3):262–278.
6. American College of Radiology. Liver Imaging Reporting and Data System version 2013.1. http://www.acr.org/Quality-Safety/Resources/LIRADS/. Published 2013. Accessed January 2013.
7. Clavien PA, Lesurtel M, Bossuyt PM, et al. Recommendations for liver transplantation for hepatocellular carcinoma: an international consensus conference report. Lancet Oncol 2012;13(1):e11–e22.
8. Mazzaferro V, Regalia E, Doci R, et al. Liver transplantation for the treatment of small hepatocellular carcinomas in patients with cirrhosis. N Engl J Med 1996;334(11):693–699.
9. Mazzaferro V, Bhoori S, Sposito C, et al. Milan criteria in liver transplantation for hepatocellular carcinoma: an evidence-based analysis of 15 years of experience. Liver Transpl 2011;17(Suppl 2):S44–S57.
10. Medical Imaging Resource Center (MIRC) Digital Imaging and Communications in Medicine. (DICOM) Anonymizer. Radiological Society of North America. http://mircwiki.rsna.org/index.php?title=The_MIRC_DICOM_Anonymizer. Accessed January 29, 2014.
11. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. Fam Med 2005;37(5):360–363.
12. American College of Radiology. Breast Imaging Reporting and Data System, 4th ed. http://www.acr.org/Quality-Safety/Resources/BIRADS/Mammography. Accessed January 29, 2014.