Title
Interreader Reliability of LI-RADS Version 2014 Algorithm and Imaging Features for Diagnosis of Hepatocellular Carcinoma: A Large International Multireader Study

Permalink
https://escholarship.org/uc/item/3892c4b7

Journal
RADIOLOGY, 286(1)

ISSN
0033-8419

Authors
Fowler, KJ
Tang, A
Santillan, C
et al.

Publication Date
2018

DOI
10.1148/radiol.2017170376

Peer reviewed
Purpose:
To determine in a large multicenter multireader setting the interreader reliability of Liver Imaging Reporting and Data System (LI-RADS) version 2014 categories, the major imaging features seen with computed tomography (CT) and magnetic resonance (MR) imaging, and the potential effect of reader demographics on agreement with a preselected nonconsecutive image set.

Materials and Methods:
Institutional review board approval was obtained, and patient consent was waived for this retrospective study. Ten image sets, comprising 38–40 unique studies (equal number of CT and MR imaging studies, uniformly distributed LI-RADS categories), were randomly allocated to readers. Images were acquired in unenhanced and standard contrast material–enhanced phases, with observation diameter and growth data provided. Readers completed a demographic survey, assigned LI-RADS version 2014 categories, and assessed major features. Intraclass correlation coefficient (ICC) assessed with mixed-model regression analyses was the metric for interreader reliability of assigning categories and major features.

Results:
A total of 113 readers evaluated 380 image sets. ICC of final LI-RADS category assignment was 0.67 (95% confidence interval [CI]: 0.61, 0.71) for CT and 0.73 (95% CI: 0.68, 0.77) for MR imaging. ICC was 0.87 (95% CI: 0.84, 0.88) for arterial phase hyperenhancement, 0.85 (95% CI: 0.81, 0.87) for washout appearance, and 0.84 (95% CI: 0.80, 0.87) for capsule appearance. ICC was not significantly affected by liver expertise, LI-RADS familiarity, or years of postresidency practice (ICC range, 0.69–0.70; ICC difference, 0.003–0.01 [95% CI: −0.003 to −0.01, 0.004–0.02]). ICC was borderline higher for private practice readers than for academic readers (ICC difference, 0.009; 95% CI: 0.000, 0.021).

Conclusion:
ICC is good for final LI-RADS categorization and high for major feature characterization, with minimal reader demographic effect. Of note, our results using selected image sets from nonconsecutive examinations are not necessarily comparable with those of prior studies that used consecutive examination series.

Kathryn J. Fowler, MD
An Tang, MD, MSc, FRCPC
Cynthia Santillan, MD
Mythreyi Bhargavan-Chatfield, PhD
Jay Heiken, MD
Reena C. Jha, MD
Jeffrey Weinreb, MD
Hero Hussain, MD
Donald G. Mitchell, MD
Mustafa R. Bashir, MD
Eduardo A. C. Costa, MD
Guilherme M. Cunha, MD
Laura Coombs, PhD
Tanya Wolfson, MA
Anthony C. Gamst, PhD
Giuseppe Brancatelli, MD
Benjamin Yeh, MD
Claude B. Sirinli, MD

1 From the Mallinckrodt Institute of Radiology, Washington University School of Medicine, 510 S Kingshighway Blvd, St Louis, MO 63110 (K.J.F., J.H.); Department of Radiology, Centre Hospitalier de l’Universite de Montreal, Montreal, Canada (A.T.); Department of Radiology, Liver Imaging Group (C.S., C.B.S.), and Computational and Applied Statistics Laboratory, San Diego Supercomputer Center (T.W., A.C.G.), University of California San Diego, San Diego, Calif; American College of Radiology, Reston, Va (M.B., L.C.); Department of Radiology, MedStar Georgetown University Hospital, Washington, DC (R.C.J.); Department of Radiology, Yale Medical School, New Haven, Conn (J.W.); Department of Radiology, University of Michigan, Ann Arbor, Mich (H.H.); Department of Radiology, Thomas Jefferson University, Philadelphia, Pa (O.G.M.); Department of Radiology, Center for Advanced Magnetic Resonance Development, Duke University Medical Center; Durham, NC (M.R.B.); Cedral, CT and MRI, Joao Pessoa, Brazil (E.A.C.C.); Clinica de Diagnostico por Imagem CDA-DASA, Rio de Janeiro, Brazil (G.M.C.); Division of Radiological Science, Di.Bi.Med., University of Palermo, Palermo, Italy (G.B.); and Department of Radiology, University of California San Francisco, San Francisco, Calif (B.Y.). Received February 15, 2017; revision requested April 10; revision received June 9; accepted July 18; final version accepted August 24. Address correspondence to K.J.F. (e-mail: fowlerk@mrv.wustl.edu).

Supported by a research scholarship from the Fonds de Recherche du Québec–Santé and Fondation de l’association des radiologistes du Québec (FRQS-ARQ #26993).

© RSNA, 2017
Established hepatocellular carcinoma (HCC) imaging features include arterial phase hyperenhancement (APHE), washout appearance, and capsule appearance. Combinations of these features with size and growth are integral to imaging algorithms (1–6), as summarized in Table 1 (7,8). Unlike other algorithms, LI-RADS provides ordinal observation categories, a standardized lexicon, and major and ancillary features of HCC (Table 2).

Although there is extensive evaluation of imaging accuracy for HCC in the literature (9), the lack of a standardized lexicon, inconsistent and ambiguous definitions for imaging features, and variable protocols challenge the synthesis of available evidence. LI-RADS promotes standardization and, consequently, reproducibility across institutions and radiologists. Few studies have evaluated LI-RADS reproducibility, each with limitations of single-center retrospective cohorts and, often, with narrow focus on binary diagnosis of HCC rather than on the whole algorithm (10–12). The interreader reliability (IRR) of individual features and their incorporation into a complex algorithm remain gaps in knowledge.

The primary aim of our study is to determine in a large multicenter multireader setting the IRR of LI-RADS categories and major imaging features seen with CT and MR imaging and to determine the potential effect of reader demographics on agreement. The secondary aim is to assess the contribution of ancillary features and tie-breaking rules toward categorization.

**Advances in Knowledge**

- In this study of selected examinations and images showing liver lesions, the overall intra-class correlation coefficient (ICC) of Liver Imaging Reporting and Data System (LI-RADS) categorization was good for CT (0.67) and MR imaging (0.73), and the inclusion or exclusion of LI-RADS category M studies and the combination of LI-RADS categories 4 and 5 versus the separation of these categories did not affect overall agreement.

- Overall ICC was excellent for arterial phase hyperenhancement (0.87), washout appearance (0.85), and capsule appearance (0.84) and was similar for all major features at both CT and MR imaging.

- ICC was not significantly affected by liver imaging expertise, a priori LI-RADS familiarity, or years of postresidency practice (ICC range: 0.69–0.70; ICC difference, 0.003–0.001); there was borderline difference in agreement between readers in academic practice and those in a private or mixed practice environment (ICC difference, 0.009).

- Ancillary features and tie-breaking rules were used for 9.8% and 8.8% of review forms completed, respectively, and were more frequently used for MR imaging than for CT.

- Our results for ICC of major features and overall categorization were higher than those in previous publications; this may be related to the limitation of preselected observations rather than consecutive observations from clinical practice.

**Materials and Methods**

**Collaboration and Logistical Support**

Logistical support for this study was provided by the American College of Radiology (ACR). Two ACR employees (M.B., L.C.) are coauthors of this article. They helped design the study, oversaw the block randomizations and case assignments, and maintained the online case review forms and reader demographic database. Coauthors who are not employees of or consultants for ACR (K.J.F., C.S., C.B.S.) had control over the inclusion of all data and information submitted for publication.

**Design**

Institutional review board approval was obtained, and patient consent was waived for this Health Insurance Portability and Accountability Act–compliant retrospective multicenter multireader study.

**Liver CT and MR Imaging Atlas**

An atlas of CT and MR imaging LI-RADS observations was constructed from multiphase images contributed by nine radiologists (J.H., C.S., D.G.M., J.W., H.H., B.Y., M.B., R.C.J., and C.B.S.; approximately 9–35 years of experience) on the LI-RADS Steering Committee. All images were acquired or reviewed for clinical care from 2008 to 2013 at the members’ institutions. Imaging studies...
Diagnostic Imaging Criteria for Definite HCC

<table>
<thead>
<tr>
<th>Diagnostic System</th>
<th>Imaging Criteria for Definite HCC</th>
<th>Specific Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>AASLD*</td>
<td>Lesion seen at antecedent US. At CT or MR imaging, lesion diameter &gt;1 cm indicates APHE and washout</td>
<td>Applied in context of US surveillance.</td>
</tr>
<tr>
<td>OPTN†</td>
<td>At CT or MR imaging, for lesions ≥1 cm and &lt;2 cm, one of the following: (a) APHE, washout, and capsule growth or (b) APHE and ≥50% growth in 6 months or less. For lesions ≥2 cm and ≤5 cm, one of the following: (a) APHE and washout, (b) APHE and capsule, or (c) APHE and threshold growth</td>
<td>Applied in context of MELD, exception point allocation for transplant candidates.</td>
</tr>
<tr>
<td>LI-RADS‡</td>
<td>For lesions 1.0 to 1.9 cm, APHE and two of the following: (a) washout, capsule growth, or threshold growth; (b) lesion seen at antecedent US and APHE and washout; or (c) APHE and ≥50% growth in 6 months or less. For lesions ≥2 cm, APHE and at least one of the following: washout, capsule growth, or threshold growth</td>
<td>Final categorization (LR-1 to LR-5, LR-5V, LR-M)</td>
</tr>
</tbody>
</table>

Note.—AASLD = American Association for the Study of Liver Disease, CT = computed tomography, LI-RADS = Liver Imaging Reporting and Data System, MELD = Model for End-Stage Liver Disease, MR = magnetic resonance, OPTN = Organ Procurement and Transplantation Network, US = ultrasonography.

* Source.—Reference 1.
† Source.—Reference 7.
‡ Source.—Reference 8.
§ Source.—Reference 1.
¶ Source.—Reference 8.

For further details and definitions regarding ancillary features and tie-breaking rules, please refer to the American College of Radiology (ACR) website: http://www.acr.org/Quality-Safety/Resources/LIRADS/.

Table 2

<table>
<thead>
<tr>
<th>LI-RADS Version 2014 Categories and Definitions</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR-1</td>
</tr>
<tr>
<td>LR-2</td>
</tr>
<tr>
<td>LR-3</td>
</tr>
<tr>
<td>LR-4</td>
</tr>
<tr>
<td>LR-5</td>
</tr>
<tr>
<td>LR-5V</td>
</tr>
<tr>
<td>LR-M</td>
</tr>
</tbody>
</table>

were selected to represent a mix of classic imaging features and more equivocal findings, with the intention of capturing a wide range of observations that reflect institutional practice. For MR imaging, only observations obtained by using an extracellular space intravenous contrast agent or gadobenate were included. The LI-RADS observations ranged from benign to malignant. For each observation, the contributing Steering Committee member reviewed the entire examination, including images at different phases of contrast enhancement. MR images depicting observations were obtained with standard liver protocol sequences (eg, T2-weighted imaging, diffusion-weighted imaging, unenhanced T1-weighted in- and opposed-phase imaging, and dynamic unenhanced and contrast-enhanced imaging in the late arterial, portal venous, and delayed phases, along with arterial subtraction imaging).

Image Sets

All observations (n = 382) within the atlas were reviewed in consensus by two members of the LI-RADS Steering Committee (K.J.F., C.S.; 3 and 10 years of experience, respectively). Each observation was assigned a LI-RADS version 2014 category in consensus. The range of observation categories in the atlas included 21 LR-1 lesions, 56 LR-2 lesions, 96 LR-3 lesions, 74 LR-4 lesions, 91 LR-5 lesions, 28 LR-5V lesions, and 16 LR-M lesions. All observations were then divided into 10 case sets, each comprising 38–40 individual observations selected via block randomization to represent an equal representation
of CT and MR imaging and an approximately even distribution of LR-1 through LR-5, LR-5V, and LR-M findings. The number of observations per case set was selected to represent a large sample of observations but not more than could be reasonably completed in one review session. The finalized case sets were displayed in a standardized format and were subsequently converted from PowerPoint to Portable Document Format for distribution to readers.

### Case Review Forms

Online case review forms were developed (https:// surveymonkey.com). Readers were required to enter the case identification number and, as quality assurance indicators, the modality (CT or MR imaging) and observation size. The review forms detailed the LI-RADS version 2014 diagnostic algorithm. Readers were asked to score the presence of major features, assign an initial LI-RADS category, score the presence of LI-RADS ancillary features, apply tie-breaking rules, and assign a final LI-RADS category. Each reader also completed a demographic survey to record their familiarity with LI-RADS, the number of years they had been in practice, their expertise in liver imaging with CT and MR imaging, and their experience with other imaging algorithms. As none of the case sets included gadoxetate-enhanced MR images or any US images, the review forms did not address imaging features specific to hepatobiliary contrast agents or the LR-5US category.

### Readers

A total of 302 radiologists were invited to participate via e-mail, and 113 agreed to participate. Potential readers were selected based on their expertise in liver imaging, their institutional affiliation with a Steering Committee member, or referral from other radiologists. The list of radiologists was meant to capture broad representation of experts and nonexperts from many institutions in both academic and community settings, including those outside the United States. Radiologists were provided a brief introduction to the study and the rationale behind it. Radiologists who confirmed their participation received internet links to downloadable case sets, online review forms, instructions to complete the review forms, and case set assignment. Case sets were assigned by block randomization to these radiologists so that each case set would be reviewed by an equal number of readers. Readers were allowed access to LI-RADS version 2014 materials available online (https://www.acr.org/Quality-Safety/Resources/LIRADS) but received no additional training.

### Statistical Analyses

Statistical analyses were performed with R software (version 3.3.1 [2016]; The R Foundation for Statistical Computing, Vienna, Austria). For Markov Chain Monte Carlo (MCMC) estimation, we used the rjags package (https://www.rjags.org/).
correlation coefficient (ICC), with 95% confidence intervals (CIs) as a measure of IRR for assigning LI-RADS categories and major features (13,14). ICC values were categorized according to Landis and Koch (15) as follows: ICC less than 0.20 indicated poor agreement; ICC of 0.21–0.40, fair agreement; ICC of 0.41–0.60, moderate agreement; ICC of 0.61–0.80, substantial agreement; ICC of 0.81–1.00, almost perfect agreement.

ICC accounts for the magnitude of disagreement between readers and, under certain conditions, is equal to the weighted k value (16). A linear mixed model (LMM) was used to model LI-RADS categories as a function of cases and readers, with case- and reader-specific intercepts fitted. ICC was computed from the case, reader, and error variances of the LI-RADS categories and the application of tie-breaking rules were summarized. Mixed-effects regression analysis was used to estimate intraclass correlation coefficients (ICCs).

### Table 4

<table>
<thead>
<tr>
<th>Finding</th>
<th>CT</th>
<th>MR Imaging</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LI-RADS category</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR-1/LR-2, LR-3, LR-4, LR-5/LR-5V, LR-M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR-1/LR-2, LR-3, LR-4, LR-5/LR-5V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR-1/LR-2, LR-3, LR-4/LR-5/LR-5V, LR-M</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LR-1/LR-2, LR-3, LR-4/LR-5/LR-5V</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Major features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>APHE</td>
<td>0.86 (0.82, 0.91)</td>
<td>0.88 (0.83, 0.94)</td>
<td>0.87 (0.84, 0.90)</td>
</tr>
<tr>
<td>Washout appearance</td>
<td>0.84 (0.80, 0.90)</td>
<td>0.84 (0.80, 0.88)</td>
<td>0.85 (0.81, 0.88)</td>
</tr>
<tr>
<td>Capsule appearance</td>
<td>0.79 (0.72, 0.85)</td>
<td>0.88 (0.83, 0.92)</td>
<td>0.84 (0.80, 0.87)</td>
</tr>
</tbody>
</table>

Note.—Data are ICCs. Data in parentheses are 95% CIs. LR-M cases were included in major features analyses.

### Figure 1

**Case 525**

**CT**

Please score the observation (45 mm) indicated by straight white arrow.

Figure 1: CT image set shows an LR-5 observation (arrows) with high agreement. The image set is exactly as it was presented to reviewers. All (100%) reviewers chose LR-5 as the category. DP = delayed phase, HAP = hepatic arterial phase, Pre = precontrast, PVP = portal venous phase.
regression model. Parametric bootstrap analysis was used to construct 95% CIs around the ICCs. For IRR of LI-RADS categories, the following four stratifications were tested: (a) LR-1/LR-2, LR-3, LR-4, LR-5/LR-5 V/LR-M; (b) LR-1/LR-2, LR-3, LR-4, LR-5/LR-5 V; (c) LR-1/LR-2, LR-3, LR-4/LR-5/LR-5 V/LR-M; and (d) LR-1/LR-2, LR-3, LR-4/LR-5/LR-5 V. In all stratifications, LR-1 and LR-2 were pooled because they represent observations that were probably or definitely benign and would require no follow-up (17). Similarly, LR-5 and LR-5 V were pooled because they represent definite HCC and would require no biopsy according to current clinical practice guidelines. In two of the stratifications, LR-4 was pooled with LR-5 and LR-5 V to form a composite category of probably or definitely HCC. Finally, because LR-M does not follow the same ordinal probability as other categories, the ICC was calculated both with and without inclusion of LR-M cases.

Similarly, a generalized LMM with case- and reader-specific intercepts was used to compute ICC for IRR on major features (binary outcomes). For these models, the estimation method was MCMC (18), and the 95% CIs were computed from the MCMC-generated distribution.

IRR was compared between participating radiologists and was based on their self-reported proficiency in liver imaging, familiarity with LI-RADS, years of posttraining practice, and overall expertise. The four characteristics were dichotomized (highly proficient vs other, very familiar with LI-RADS vs other, etc) and were examined in separate analyses. LMM and generalized LMMs were extended with the addition of reader group strata, which enabled computation of separate ICCs (eg, for experts vs nonexperts) within one model. The significance of the difference between reader subgroup ICCs was assessed by using parametric bootstrap analysis in the LMM and MCMC-generated distribution in the generalized LMM. LMM together with a parametric (model-based) bootstrap for 95% CI construction around the ICC and ICC differences were used. The LMM was refit 1000 times by using pseudodata samples. ICCs and their differences were computed at each iteration. The 95% CIs around each ICC and their difference were constructed from the bootstrap replicates. Two-tailed $P < .05$ was considered indicative of a significant difference. For these analyses, CT and MR imaging were combined, LR-M was included, and only the LR-1/LR-2, LR-3, LR-4, LR-5/LR-5 V, LR-M stratification was used to assess agreement on ordinal LI-RADS categories. The $P$ values derived from the analysis for categories and major features were not adjusted for multiple
Readers

Of the 302 readers who were invited, 167 (55%) confirmed they would participate. Of these 167 who were sent demographic forms with links for case assignment, 54 (32%) dropped out, while the remaining 113 (68%) completed the case review forms. Table 3 summarizes demographics of the 113 final readers, who completed a total of 4346 case review forms. In instances in which duplicate case review forms were completed by the same reader, the second form was included in analysis under the assumption that the first form contained errors the reader intended to correct. Three cases and all corresponding review forms were excluded from final analysis because of incorrect labeling of contrast enhancement phases that was noticed only after case distribution. The final data set included 380 unique cases and 4009 separate review forms. Each case set was reviewed by an average of 11 readers (range, six to 17 readers). The unequal distribution of case sets was due to reader drop out and some readers downloading and scoring the wrong case sets. Readers were asked to input the case number for each form to ensure the appropriate case was scored.

Agreement on LI-RADS Categories

The agreement for LI-RADS categories was analyzed according to four distinct groups and is shown in Table 4. The ICC for agreement was highest for MR imaging for all groups, and agreement was slightly higher for variants without the combined LR-4/LR-5 category. The exclusion of LR-M had little effect on ICC. Figures 1–4 show examples of cases with high and low agreement. The image sets are presented as they were presented to reviewers. Consensus reading by the steering committee was LR-M. The most frequent category assigned by reviewers was LR-5, assigned by approximately 33% of reviewers. This image set is a challenging example, and the lesion (arrows) could arguably be categorized as LR-4. DP = delayed phase, HAP = hepatic arterial phase, Pre = precontrast, PVP = portal venous phase.
low agreement at both CT and MR imaging. Figures 5 and 6 are graphs of average score distribution for all cases at CT and MR imaging. Reader agreement was relatively high, with few outliers in each category and similar results between CT and MR imaging.

**Figure 4**

![MR image set with low agreement](image)

**Case 529 3T MRI**

<table>
<thead>
<tr>
<th></th>
<th>Pre</th>
<th>HAP</th>
<th>Delay = 2 min</th>
<th>Delay = 3 min</th>
<th>Delay = 4 min</th>
<th>Delay = 15 min</th>
</tr>
</thead>
<tbody>
<tr>
<td>OP, TE</td>
<td>1.1 ms</td>
<td>IP, TE = 2.3 ms</td>
<td>T2w GRE</td>
<td>T2w SSFSE</td>
<td>DWI, $b = 800$</td>
<td>R2* map</td>
</tr>
</tbody>
</table>

**Figure 4:** MR image set with low agreement. The image set is exactly as it was presented to reviewers. Although the most frequent response was LR-5, only 36% of reviewers scored this lesion (arrows) as LR-5. DWI = diffusion-weighted imaging, GRE = gradient-recalled echo, HAP = hepatic arterial phase, IP = in phase, OP = opposed phase, Pre = precontrast, SSFSE = single-shot fast spin-echo, TE = echo time, T2W = T2-weighted.

**Agreement and Reader Demographics**

ICCs for assigning major features are shown in Table 4. Overall ICC was high for APHE (0.87; 95% CI: 0.84, 0.90), washout appearance (0.85; 95% CI: 0.81, 0.88), and capsule appearance (0.84; 95% CI: 0.80, 0.87). ICC was slightly higher with MR imaging than with CT for all features. Agreement was not analyzed for diameter or growth, which were provided to the readers. On the basis of the majority opinion, 50 lesions were smaller than 10 mm (CT, $n = 16$; MR imaging, $n = 34$), 126 were 10–19 mm (CT, $n = 72$; MR imaging, $n = 54$), and 204 were 20 mm or larger (CT, $n = 104$; MR imaging, $n = 100$).

**Ancillary Features**

Ancillary features were used in 10% (393 of 4009) of review forms to modify the final category. Figure 7 shows the relative frequency of ancillary features in cases that were upgraded or downgraded by one category. Ancillary features were present more commonly on MR images than on CT images (74% [291 of 393] vs 26% [102 of 393]). The individual effect of each ancillary feature is difficult to discern because some cases that were upgraded or downgraded had multiple features selected. Although review forms asked specifically if ancillary features were used to determine the final category, they did require specification of which ones were used in the case of multiple selections.

**Discussion**

LI-RADS showed good ICC (range, 0.70–0.71) for LR-1 through LR-5 and LR-M category observations, with similar reliability shown for CT and MR imaging. The exclusion of LR-M findings did not meaningfully affect
overall reliability. ICC incorporates the magnitude of difference between readers, allowing for a global assessment of larger versus smaller disagreements between readers (i.e., a disagreement between LR-1 and LR-5 categories has a greater effect on ICC than does a disagreement between LR-1 and LR-2 categories).

Six prior studies have assessed reader agreement regarding LI-RADS categories (11,19,20) or major features of HCC (11,19–22). Unlike these studies, which all included readers from the same institution, ours is further enhanced by (a) multicenter international cross-sectional reader pool, including community practice, academic, and mixed practice environments; (b) derivation of the cases from eight different sites; (c) the largest number of readers tested to date; (d) no training module aside from the LI-RADS materials available online; (e) mixture of all LI-RADS categories, including LR-M, at CT and MR imaging; (f) reduction of variability related to workstation adjustments (windowing, mislabeling of lesions, etc); (g) no requirement for pathologic proof of diagnosis, removing potential verification bias; and (h) thorough evaluation of LI-RADS, including data on use of ancillary features and tie-breaking rules. Table 6 compares the results of prior studies with ours.

The ICC for LI-RADS categories in our study (i.e., 0.65–0.71) is similar to that reported in prior studies (i.e., 0.44–0.82). We believe that our results accurately reflect reproducibility, given our large pool of cases (380 total, almost twice that of any other study) and our large pool of readers. These help control for problematic cases and equivocal features, which may negatively affect ICC.

Our major feature agreement was higher than that in previous publications (11,24). In particular, the ICC for capsule appearance in our study, 0.80 (95% CI: 0.74, 0.87) for CT and 0.89 (95% CI: 0.84, 0.94) for MR imaging, is higher than that reported in prior publications (0.37 and 0.65, respectively), with higher agreement for MR imaging in studies that evaluated both MR imaging and CT. Despite the intention to create an atlas with a mix of cases, the selection of cases may have favored those with more classic features and contributed to the higher ICC in our study. The lower agreement for “capsule” reported in the literature may be partly due to the difficulty in distinguishing a distinct delayed enhancing rim from the background fibrosis, which can be even more challenging in smaller lesions (25–27). Washout appearance agreement values were 0.85 (95% CI: 0.81, 0.88) for CT and 0.84 (95% CI: 0.79, 0.88) for MR imaging in our study; these values were greater than the 0.48–0.72 range reported in the literature. The subjective assessment of washout appearance may be confounded by background liver alterations (28), and some authors have suggested a quantitative approach to defining washout (29). The higher ICC results for major features in our study may underscore the variability that can be generated from workstation adjustments, which were removed in our study design. This raises the possibility that standardizing workstation adjustments may be a method to improve radiologist agreement, and this topic merits further investigation. Likewise, the selection of images potentially showing “classic” features may have contributed to higher ICC in our study. Additionally, the reviewers were forced to use the algorithm in a stepwise fashion, which may have improved reproducibility. Reader demographics minimally affected ICC, indicating LI-RADS can be applied reproducibly by readers of varying LI-RADS familiarity, liver expertise, or practice setting.

Ancillary features were frequently observed but resulted in a change in final category in only approximately
10% of total cases. Tie-breaking rules were used in a minority (9% of case review forms). Ancillary features and the use of tie-breaking rules were more prevalent for MR imaging, likely owing to a wider variety of image contrasts, the greater number of sequences, and the larger number of potential ancillary features available. The minor role of ancillary features in our study is similar to others’ experience, suggesting that they may have a minimal effect on final diagnosis or reproducibility (30). The frequency and specificity of ancillary features requires further investigation in clinical trials to help determine their optimal incorporation in the imaging algorithm. The effect of individual ancillary features and tie-breaking rules on interreader agreement was not assessed with our methods, but it should be the topic of future study.

Our study had several limitations. To achieve a large interrater agreement study, case distribution was limited to prepared cases with select images and annotations. This construct may not be optimal for assessing all imaging features, many of which may require the reviewer to scroll or alter window settings to be

---

**Figure 6**: Graph shows reader agreement, with color coding of LI-RADS categories to show the distribution and average of scores for each individual image set at MR imaging. Consensus read refers to the consensus reading of the steering committee. Each point represents the average score for a given observation. Bars represent the standard deviation.

**Table 5**

<table>
<thead>
<tr>
<th>Practice Type</th>
<th>Overall LI-RADS Category</th>
<th>APHE</th>
<th>Capsule</th>
<th>Washout Appearance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Academic</td>
<td>0.71 (0.67, 0.73)</td>
<td>0.86 (0.82, 0.90)</td>
<td>0.84 (0.80, 0.89)</td>
<td>0.85 (0.81, 0.89)</td>
</tr>
<tr>
<td>Community or mixed</td>
<td>0.69 (0.66, 0.72)</td>
<td>0.88 (0.82, 0.93)</td>
<td>0.83 (0.77, 0.89)</td>
<td>0.83 (0.77, 0.89)</td>
</tr>
<tr>
<td>P-value</td>
<td>.500*</td>
<td>.613</td>
<td>.783</td>
<td>.880</td>
</tr>
<tr>
<td>Liver expertise</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Yes</td>
<td>0.70 (0.66, 0.73)</td>
<td>0.88 (0.84, 0.91)</td>
<td>0.83 (0.79, 0.87)</td>
<td>0.86 (0.82, 0.89)</td>
</tr>
<tr>
<td>Some or none</td>
<td>0.70 (0.66, 0.73)</td>
<td>0.81 (0.69, 0.89)</td>
<td>0.94 (0.74, 1.0)</td>
<td>0.75 (0.63, 0.85)</td>
</tr>
<tr>
<td>P-value</td>
<td>.826</td>
<td>.134</td>
<td>.248</td>
<td>.0392*</td>
</tr>
<tr>
<td>LI-RADS familiarity</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Very familiar</td>
<td>0.70 (0.67, 0.74)</td>
<td>0.88 (0.81, 0.97)</td>
<td>0.83 (.76, 0.89)</td>
<td>0.92 (0.86, 0.97)</td>
</tr>
<tr>
<td>Somewhat or not familiar</td>
<td>0.70 (0.67, 0.73)</td>
<td>0.87 (0.83, 0.90)</td>
<td>0.84 (0.79, 0.89)</td>
<td>0.82 (0.78, 0.86)</td>
</tr>
<tr>
<td>P-value</td>
<td>.264</td>
<td>.820</td>
<td>.707</td>
<td>.0072*</td>
</tr>
<tr>
<td>Years of practice</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>≥6 years</td>
<td>0.70 (0.66, 0.73)</td>
<td>0.86 (0.83, 0.90)</td>
<td>0.84 (0.80, 0.88)</td>
<td>0.84 (0.80, 0.87)</td>
</tr>
<tr>
<td>&lt;6 years</td>
<td>0.71 (0.67, 0.74)</td>
<td>0.93 (0.83, 0.99)</td>
<td>0.85 (0.72, 0.98)</td>
<td>0.91 (0.81, 0.98)</td>
</tr>
<tr>
<td>P-value</td>
<td>.852</td>
<td>.198</td>
<td>.911</td>
<td>.204</td>
</tr>
</tbody>
</table>

Note.—Bootstrap-based (for LI-RADS categories) and MCMC-based (for major features) P-values for the differences in ICC between strata are shown. Data in parentheses are 95% CIs. * P-values indicate a significant difference. These P-values are presented without an adjustment for multiple comparisons to highlight potential differences for further exploration.
In conclusion, in this multicenter study of interrater agreement on HCC imaging diagnosis, the overall agreement was good for final LI-RADS categorization and high for major feature characterization, with minimal reader demographic effects.

Our results for agreement for major features are higher than those in previous consecutive case series, possibly because variability from display setting adjustment was eliminated and because selected annotated images were assessed. Reader demographics showed little effect on ICC. Ancillary features and tie-breaking rules were used in a minority of cases and were more frequently used for MR imaging than for CT. Future studies are required to determine the optimal role of ancillary features for categorization.

Disclosures of Conflicts of Interest: K.J.F. disclosed no relevant relationships. A.T. disclosed no relevant relationships. C.S. Activities related to the present article: disclosed no relevant relationships. A.G. Activities related to the present article: disclosed no relevant relationships. E.A.C.C. Activities related to the present article: disclosed no relevant relationships. T.W. Activities related to the present article: disclosed no relevant relationships. A.G. Activities related to the present article: disclosed no relevant relationships. R.C.I. Activities related to the present article: disclosed no relevant relationships. J.H. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. B.R. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. H.H. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. D.G.M. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. M.R.B. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. J.W. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. L.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. G.M.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. E.A.C.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. G.B. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. B.Y. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. C.S. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. R.C.I. Activities related to the present article: disclosed no relevant relationships. D.G.M. Activities related to the present article: disclosed no relevant relationships. A.G. Activities related to the present article: disclosed no relevant relationships. E.A.C.C. Activities related to the present article: disclosed no relevant relationships. T.W. Activities related to the present article: disclosed no relevant relationships. A.G. Activities related to the present article: disclosed no relevant relationships. R.C.I. Activities related to the present article: disclosed no relevant relationships. J.H. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. H.H. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. D.G.M. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. M.R.B. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. J.W. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. L.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. G.M.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. E.A.C.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. G.B. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. B.Y. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. C.S. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. R.C.I. Activities related to the present article: disclosed no relevant relationships. D.G.M. Activities related to the present article: disclosed no relevant relationships. A.G. Activities related to the present article: disclosed no relevant relationships. E.A.C.C. Activities related to the present article: disclosed no relevant relationships. T.W. Activities related to the present article: disclosed no relevant relationships. A.G. Activities related to the present article: disclosed no relevant relationships. R.C.I. Activities related to the present article: disclosed no relevant relationships. J.H. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. H.H. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. D.G.M. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. M.R.B. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. J.W. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. L.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. G.M.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. E.A.C.C. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. G.B. Activities related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. B.Y. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. Other relationships: disclosed no relevant relationships. C.S. Activities related to the present article: disclosed no relevant relationships.
### Table 6

**Summary of Agreement from Literature**

<table>
<thead>
<tr>
<th>Study and LI-RADS Version</th>
<th>No. of Readers, No. of Lesions, and Case Selection Criteria</th>
<th>Modality and Format of Image Sets</th>
<th>Feature*</th>
<th>LI-RADS Category*</th>
<th>AF Use (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Davenport M (LI-RADS version 2013.1)¹</td>
<td>10 readers (5 expert, 5 novice), 100 lesions, consecutive cases scored by research team from 2011–2013 to meet quota of 20 per LI-RADS category</td>
<td>MR imaging ECA phases, APHE, ( k = 0.67 ) (95% CI: 0.65, 0.69); washout, ( k = 0.48 ) (95% CI: 0.46, 0.5); capsule, ( k = 0.52 ) (95% CI: 0.5, 0.54); diameter, ICC = 0.97 (95% CI: 0.96, 0.97)</td>
<td></td>
<td>LR-1, ( k = 0.54 ) (95% CI: 0.51, 0.57); LR-2, ( k = 0.11 ) (95% CI: 0.08, 0.14); LR-3, ( k = 0.26 ) (95% CI: 0.23, 0.28); LR-4, ( k = 0.28 ) (95% CI: 0.25, 0.31); LR-5, ( k = 0.62 ) (95% CI: 0.59, 0.65)</td>
<td>22</td>
</tr>
<tr>
<td>Barth et al (LI-RADS version 2013.1)²</td>
<td>Four readers (8–12 years of experience), 104 lesions, consecutive cases from database of LI-RADS categories screened for quota of about 15 per LI-RADS category</td>
<td>MR imaging with ECA, scrollable image sets, and a screenshot showing observation</td>
<td>APHE, ( k = 0.51 ) (95% CI: 0.38, 0.65); washout, ( k = 0.52 ) (95% CI: 0.39, 0.65); capsule, ( k = 0.37 ) (95% CI: 0.23, 0.52); diameter, ( k = 0.97 ) (95% CI: 0.965, 0.98)</td>
<td>All LI-RADS categories, ( k = 0.44 ) (95% CI: 0.37, 0.52)</td>
<td>8</td>
</tr>
<tr>
<td>Zhang et al (LI-RADS version not specified)³</td>
<td>Two readers (6 years of postfellowship experience), 118 lesions, consecutive cases from existing database selected to include cases with both CT and MR images available</td>
<td>MR imaging with ECA, CT, and scrollable image sets</td>
<td>For CT: diameter, ICC = 0.95 (95% CI: 0.94, 0.97); APHE, ICC = 0.54 (95% CI: 0.44, 0.64); washout, ICC = 0.61 (95% CI: 0.46, 0.76); capsule, ICC = 0.56 (95% CI: 0.52, 0.6); For MR imaging: diameter, ICC = 0.98 (95% CI: 0.97, 0.99); APHE, ICC = 0.63 (95% CI: 0.54, 0.72); washout, ICC = 0.72 (95% CI: 0.64, 0.8); capsule, ICC = 0.65 (95% CI: 0.6, 0.7)</td>
<td>Not provided</td>
<td>NA</td>
</tr>
<tr>
<td>Ehman et al (LI-RADS version 2014)⁴</td>
<td>Two readers (experience not specified), 184 lesions, consecutive cases with histopathology proof of diagnosis of HCC</td>
<td>MR imaging with ECA, CT, scrollable image sets</td>
<td>For CT: APHE, ( k = 0.74 ); washout, ( k = 0.64 ); capsule, ( k = 0.46 ). For MR imaging: APHE, ( k = 0.22 ); washout, ( k = 0.58 ); capsule, ( k = 0.62 )</td>
<td>ICC for all readers both CT and MR imaging: All LI-RADS, ICC = 0.59 (95% CI: 0.53, 0.65)</td>
<td>NA</td>
</tr>
<tr>
<td>Sofue et al (LI-RADS version 2013.1)⁵</td>
<td>Two readers (7 and 6 years of experience), 200 lesions, consecutive cases with hypervascular nodules by report</td>
<td>MR imaging with EC phases, scrollable image sets</td>
<td>Washout, ( k = 0.75 ); capsule, ( k = 0.70 )</td>
<td>Clinical and research reading: all LI-RADS categories, ( k = 0.82 )</td>
<td>NA</td>
</tr>
<tr>
<td>Bashir et al (LI-RADS version 2013.1)⁶</td>
<td>Three readers (2, 4, and 12 years of experience), 200 lesions, consecutive cases with hypervascular nodules by report</td>
<td>MR imaging with ECA, CT, MR imaging with HBA, scrollable image sets</td>
<td>Washout, ( k = 0.69 ) (95% CI: 0.61, 0.77); capsule, ( k = 0.59 ) (95% CI: 0.51, 0.67); threshold growth, ( k = 0.67 ) (95% CI: 0.59, 0.75); diameter, ( k = 0.99 ) (95% CI: 0.98, 0.99)</td>
<td>LR-5 or &lt;LR-5, ( k = 0.68 ) (95% CI: 0.60, 0.76)</td>
<td>NA</td>
</tr>
<tr>
<td>Current study (LI-RADS version 2014)</td>
<td>113 readers (average 11 readers, diverse background), 380 lesions, selected cases from multiple institutions submitted to a liver atlas</td>
<td>MR imaging with ECA, CT, PowerPoint image captures</td>
<td>All CT APHE, ICC = 0.86 (95% CI: 0.82, 0.91); washout, ICC = 0.84 (95% CI: 0.80, 0.90); capsule, ICC = 0.80 (95% CI: 0.74, 0.87). All MR imaging: APHE, ICC = 0.91 (95% CI: 0.84, 0.96); washout, ICC = 0.83 (95% CI: 0.79, 0.88); capsule, ICC = 0.89 (95% CI: 0.84, 0.94).</td>
<td>ICC for all readers: MR imaging, 0.73 (95% CI: 0.68, 0.77); CT, 0.67 (95% CI: 0.61, 0.71)</td>
<td>9.8</td>
</tr>
</tbody>
</table>

Note.—A = arterial phase hyperenhancement, AF = ancillary features, C = capsule, D = diameter, ECA = extracellular agents, HBA = hepatobiliary agents, NA = not applicable, TG = threshold growth, W = washout.

* Data in parentheses are 95% CIs.

1 Source.—Reference 11.
2 Source.—Reference 19.
3 Source.—Reference 23.
4 Source.—Reference 20.
5 Source.—Reference 21.
6 Source.—Reference 22.
References


