Radiology

ORIGINAL RESEARCH • GASTROINTESTINAL IMAGING

A Multicenter Assessment of Interreader Reliability of LI-RADS Version 2018 for MRI and CT

Cheng William Hong, MD, MS • Victoria Chernyak, MD, MS • Jin-Young Choi, MD • Sonia Lee, MD • Chetan Potu, BS • Timoteo Delgado, BS • Tanya Wolfson, MA • Anthony Gamst, PhD • Jason Birnbaum, MD • Rony Kampalath, MD • Chandana Lall, MD • James T. Lee, MD • Joseph W. Owen, MD • Diego A. Aguirre, MD • Mishal Mendiratta-Lala, MD • Matthew S. Davenport, MD • William Masch, MD • Alexandra Roudenko, MD • Sara C. Lewis, MD • Andrea Siobhan Kierans, MD • Elizabeth M. Hecht, MD • Mustafa R. Bashir, MD • Giuseppe Brancatelli, MD • Michael L. Douek, MD • Michael A. Ohliger, MD PhD • An Tang, MD MSc • Milena Cerny, MD • Alice Fung, MD • Eduardo A. Costa, MD • Michael T. Corwin, MD • John P. McGahan, MD • Bobby Kalb, MD • Khaled M. Elsayes, MD • Venkateswar R. Surabhi, MD • Katherine Blair, MD • Robert M. Marks, MD • Natally Horvat, MD, PhD • Shaun Best, MD • Ryan Ash, MD • Karthik Ganesan, MD • Christopher R. Kagay, MD • Avinash Kambadakone, MD • Jin Wang, MD • Irene Cruite, MD • Bijan Bijan, MD • Mark Goodwin, MD • Guilherme Moura Cunha, MD • Dorathy Tamayo-Murillo, MD • Kathryn J. Fowler, MD • Claude B. Sirlin, MD

From the Department of Radiology and Biomedical Imaging, University of California San Francisco, 513 Parnassus Ave, S255, Box 0628, San Francisco, CA 94143 (C.W.H., M.A.O.); Liver Imaging Group, Department of Radiology, University of California, San Diego, San Diego, Calif (C.W.H., C.P., T.D., D.T.M., K.J.F., C.B.S.); Department of Radiology, Memorial Sloan Kettering Cancer Center, New York, NY (V.C., N.H.); Department of Radiology, Yonsei University, Seoul, South Korea (J.Y.C.); Department of Radiology, University of California Irvine, Orange, Calif (S.L., R.K.); Computational and Applied Statistics Laboratory, University of California San Diego, San Diego, Calif (T.W., A.G.); Department of Radiology, New York University, New York, NY (J.B.); Department of Radiology, University of Florida, Jacksonville, Fla (C.L.); Department of Radiology, University of Kentucky, Lexington, Ky (J.T.L., J.W.O.); Department of Radiology, Fundación Santa Fe de Bogotá, Bogotá, Colombia (D.A.A.); Department of Radiology, University of Michigan, Ann Arbor, Mich (M.M.L., M.S.D., W.M.); Department of Radiology, Allegheny Health Network, Pittsburgh, Pa (A.R.); Department of Radiology, Icahn School of Medicine at Mount Sinai, New York, NY (S.C.L.); Department of Radiology, New York-Presbyterian/Weill Cornell Medical Center, New York, NY (A.S.K., E.M.H.); Departments of Radiology and Medicine, Duke University Medical Center, New York, NY (M.R.B.); Section of Radiology, Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University Hospital Paolo Giaccone, Palermo, Italy (G.B.); Department of Radiology, University of California Los Angeles, Los Angeles, Calif (M.L.D.); Department of Radiology, Radiation Oncology and Nuclear Medicine, Université de Montréal, Montréal, Canada (A.T., M.C.); Department of Radiology, Oregon Health & Science University, Portland, Ore (A.F.); CEDRUL-Centro de Diagnóstico por Imagem, João Pessoa, Brazil (E.A.C.); Department of Radiology, University of California Davis, Sacramento, Calif (M.T.C., J.P.M.); Radiology Limited, Tucson, Ariz (B.K.); Department of Abdominal Imaging, University of Texas MD Anderson Cancer Center, Houston, Tex (K.M.E., V.R.S., K.B.); Department of Radiology, Naval Medical Center San Diego, San Diego, Calif (R.M.M.); University of São Paulo/Hospital Sírio-Libanês, São Paulo, Brazil (N.H.); Department of Radiology, University of Kansas, Kansas City, Kan (S.B., R.A.); Sir H. N. Reliance Foundation Hospital and Research Centre, Mumbai, India (K.G.); Department of Radiology, California Pacific Medical Center, San Francisco, Calif (C.R.K.); Department of Radiology, Massachusetts General Hospital, Boston, Mass (A.K.); The 3rd Affiliated Hospital, Sun Yat-sen University, Guangzhou, China (J.W.); Inland Imaging, Spokane, Wash (I.C.); Sutter Medical Group, Sacramento, Calif (B.B.); Austin Health, Melbourne, Australia (M.G.); Department of Radiology, University of Washington, Seattle, Wash (G.M.C.). Received November 5, 2022; revision requested December 22; revision received April 6, 2023; accepted April 28. **Address correspondence to** C.W.H. (email: *cheng.hong@ucsf.edu*).

Supported by 2017 RSNA Resident Research Grant (RR1726). Study supported by National Institutes of Health (T32 EB005970-09). A.T. supported by a Clinical Research Scholarship–Senior Salary Award by the Fonds de recherche du Québec en Santé and Fondation de l'Association des Radiologistes du Québec (FRQS-ARQ 298509).

R.M.M. is a military service member. This work was prepared as part of his official duties. Title 17, USC, § 105 provides that 'Copyright protection under this title is not available for any work of the United States Government.' Title 17, USC, § 101 defines a U.S. Government work as a work prepared by a military service member or employee of the U.S. Government as part of that person's official duties. The views expressed in this article are those of the author and do not necessarily reflect the official policy or position of the Department of the Navy, Department of Defense, or the U.S. Government.

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

See also the editorials by Johnson and Galgano and Smith in this issue.

Radiology 2023; 307(5):e222855 • https://doi.org/10.1148/radiol.222855 • Content code: GI

Background: Various limitations have impacted research evaluating reader agreement for Liver Imaging Reporting and Data System (LI-RADS).

Purpose: To assess reader agreement of LI-RADS in an international multicenter multireader setting using scrollable images.

Materials and Methods: This retrospective study used deidentified clinical multiphase CT and MRI and reports with at least one untreated observation from six institutions and three countries; only qualifying examinations were submitted. Examination dates were October 2017 to August 2018 at the coordinating center. One untreated observation per examination was randomly selected using observation identifiers, and its clinically assigned features were extracted from the report. The corresponding LI-RADS version 2018 category was computed as a rescored clinical read. Each examination was randomly assigned to two of 43 research readers who independently scored the observation. Agreement for an ordinal modified four-category LI-RADS scale (LR-1, definitely benign; LR-2, probably benign; LR-3, intermediate probability of malignancy; LR-4, probably hepatocellular carcinoma [HCC]; LR-5, definitely HCC; LR-M, probably malignant but not HCC specific; and LR-TIV, tumor in vein) was computed using intraclass correlation coefficients (ICCs). Agreement was also computed for dichotomized malignancy (LR-4, LR-5, LR-M, and LR-TIV), LR-5, and LR-M. Agreement was compared between research-versus-research reads and research-versus-clinical reads.

Results: The study population consisted of 484 patients (mean age, 62 years ± 10 [SD]; 156 women; 93 CT examinations, 391 MRI examinations). ICCs for ordinal LI-RADS, dichotomized malignancy, LR-5, and LR-M were 0.68 (95% CI: 0.61, 0.73), 0.63 (95% CI: 0.55, 0.70), 0.58 (95% CI: 0.50, 0.66), and 0.46 (95% CI: 0.31, 0.61) respectively. Research-versus-research reader agreement was higher than research-versus-clinical agreement for modified four-category LI-RADS (ICC, 0.68 vs 0.62, respectively; $P = .03$) and for dichotomized malignancy (ICC, 0.63 vs 0.53, respectively; $P = .005$), but not for LR-5 ($P = .14$) or LR-M ($P = .94$).

Conclusion: There was moderate agreement for LI-RADS version 2018 overall. For some comparisons, research-versus-research reader agreement was higher than research-versus-clinical reader agreement, indicating differences between the clinical and research environments that warrant further study.

Supplemental material is available for this article.

An earlier incorrect version appeared online. This article was corrected on June 28, 2023.

Abbreviations

HCC = hepatocellular carcinoma, ICC = intraclass correlation coefficient, LI-RADS = Liver Imaging Reporting and Data System

Summary

In an international multicenter reader study with scrollable images, overall moderate reader agreement was observed for the 2018 version of the Liver Imaging Reporting and Data System.

Key Results

- In this retrospective study of 484 patients, the Liver Imaging Reporting and Data System version 2018 was assessed using a modified four-category ordinal scale and had moderate reader agreement (ICC, 0.68).
- Binary agreement for probably or definitely malignant categories (ICC, 0.63) and LR-5 (ICC, 0.58) was moderate, whereas agreement for LR-M was poor (ICC, 0.46).
- Research-versus-research agreement differed from research-versus-clinical agreement (ICC, 0.68 vs 0.62, respectively; $P = .03$), indicating differences between these environments that warrant further study.

Multiphase CT and MRI are instrumental in the noninvasive diagnosis and management of hepatic malignancies (1). The American College of Radiology Liver Imaging Reporting and Data System (LI-RADS) standardizes the terminology, technique, interpretation, and reporting of liver imaging (2,3). LI-RADS categorizes observations from LR-1 (definitely benign) to LR-5 (definitely hepatocellular carcinoma [HCC]) and also includes categories for malignant observations without characteristic HCC features.

Higher LI-RADS categories correspond to an increasing probability of HCC (4–10). In addition, higher LI-RADS categories also have an increasing probability of progression to HCC or other malignancy at follow-up imaging (11,12). The LR-5 category, intended to be diagnostic for HCC, has an estimated specificity of 89%–99% (10,13–16). Although determining accuracy is necessary, determining precision, including reader reliability, is also necessary. Previous studies (16–24) found moderate agreement for LI-RADS but were limited by factors such as small, single-center, single-modality image sets, and/or the use of a small number of readers from a single center. A recent meta-analysis by Kang et al (25) found moderate agreement ($\kappa = 0.70$) for LI-RADS categorization but only included MRI, and 14 of the 15 studies had single-center readers. A multicenter study by Fowler et al (26) had multiple contributing sites and many readers and found moderate agreement (intraclass correlation coefficient [ICC], 0.67) for LI-RADS categorization; however, preselected image sets were used instead of fully scrollable examinations, which may overestimate reader agreement.

Previous studies assessed research reads and, to our knowledge, no prior study has incorporated reads performed in a clinical environment. Research readers, aware that their readings will be analyzed, may review cases and follow the LI-RADS algorithm more carefully. They can also read images in a controlled environment with fewer distractions. However, research readers cannot access clinical information or prior imaging and reports, and they are unable to discuss cases with other radiologists or referring clinicians. Because of these factors, studies that focus exclusively on research settings may not be fully generalizable to the clinical setting. Knowledge about LI-RADS performance in the clinical setting, including assessment of reliability or agreement with clinical reports, is needed but not yet available.

This study aims to assess reader agreement of LI-RADS in a large, international, multicenter, multireader setting using scrollable examinations. This study also incorporates deidentified clinical interpretations to gain insight into reader agreement in the clinical setting.

Materials and Methods

Study Design

This was a retrospective, multicenter, international reader study of clinically acquired multiphase LI-RADS CT and MRI examinations (Fig 1). The study was Health Insurance Portability and Accountability Act–compliant and approved by local institutional review boards, with waivers of informed consent because the research was minimal risk. Six institutions from three countries (United States, South Korea, and Colombia) submitted to the coordinating center (University of California San Diego; San Diego, Calif) deidentified examinations and reports from unique patients with at least one untreated observation. Examinations were uploaded to a cloud-based platform and assigned to two of 43 readers (Table S1), randomized so that both readers were from separate institutions that were different from the submitting site, eliminating the possibility of familiarity bias. Twenty percent of the examinations were randomly selected to be read twice by one of the readers to assess intrareader agreement.

The modalities, contrast agents, and scanner vendors at each site are in Tables S2 and S3.

Examination Selection

Submitting sites identified examinations that contained at least one untreated observation. Exclusion criteria were not applicable because only qualifying examinations were submitted. Whereas there may have been minor differences in imaging protocols, all examinations adhered to LI-RADS technical recommendations and were reported in accordance with LI-RADS reporting requirements. The examination dates ranged from October 2017 to August 2018 at the coordinating center, and dates from other sites were deidentified. For each reported observation, the report provided a unique numeric identifier, a series image number, an assigned category, and its major and ancillary features. For reports issued clinically in a language other than English, the submitting radiologist translated the report into English.

For each examination, an image analyst (C.P., with 1 year of experience) at the coordinating center reviewed the deidentified report and selected one untreated observation using a random number generator to pick among the identifiers. The corresponding observation was electronically labeled with an arrow. This was applied to the reported series and image number unless a different image identified the selected observation more clearly. The image analyst was a researcher with study-specific training for his tasks and software usage and was supervised by two authors (C.W.H., a radiology resident, and C.B.S., with >20 years of experience).

Research Reads

Labeled and deidentified examinations were uploaded to a cloud-based platform (Arterys; Arterys). The platform provided standard capabilities including scrolling, panning, magnifying, window-level adjustment, regions of interest, and measurement calipers. The readers were mostly subspecialty abdominal radiologists. Each reader scored the annotated observation using a standardized Research Electronic Data Capture form that included fields for the category and individual imaging features. Readers were given a stepwise guide for the reading platform and case report forms. Readers could review reference materials, but no training was provided, reflecting clinical practice. The research reads were performed between May 2019 and October 2020.

Each reader completed a research reader questionnaire regarding their geographic region, institutional affiliation, fellowship training, experience, familiarity with LI-RADS, and institutional practice patterns.

Clinical Reads

Deidentified reports were parsed automatically with custom software scripts (Python; Python Software Foundation), which extracted feature-level information. Reports were randomly selected for manual verification, and the features were extracted manually in cases where the clinical reports were incorrectly formatted. The reported imaging features were used to recalculate the corresponding LI-RADS version 2018 category and exclude features that require comparison to prior examinations as a rescored "clinical read" (hereafter, referred to as clinical reads). This was necessary because, whereas most of the examinations were clinically reported using LI-RADS version 2017, the research readers would apply LI-RADS version 2018 and would not have access to previous examinations.

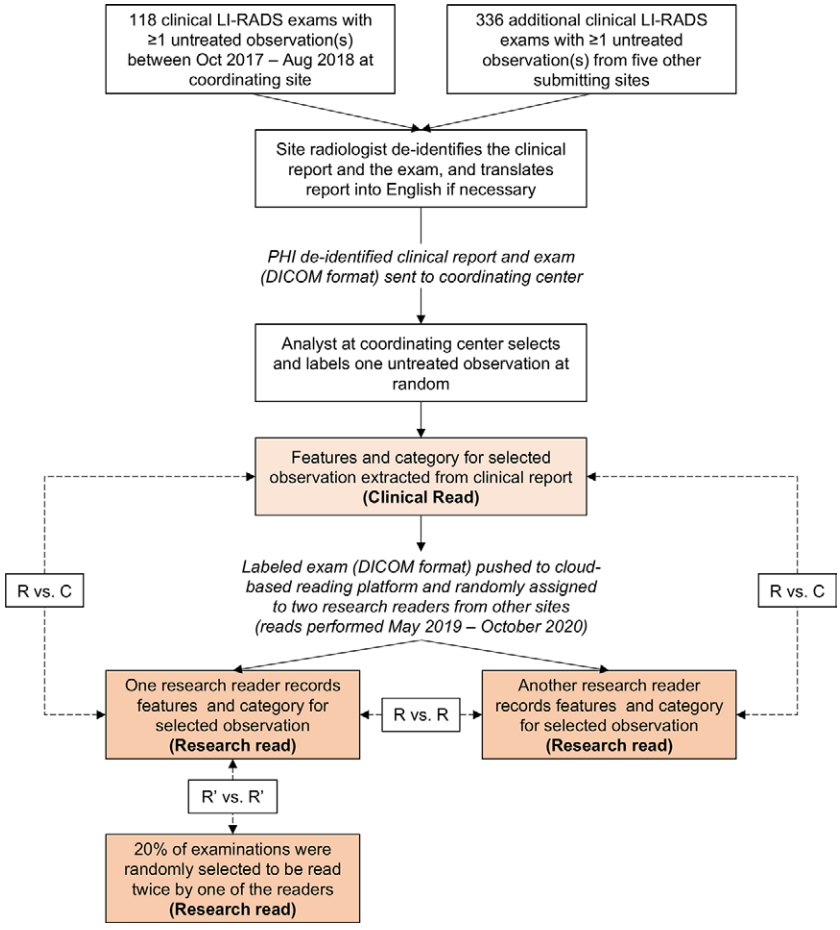

Figure 1: Schematic of the retrospective study design. Deidentified examinations from the coordinating site and five other submitting sites were randomly assigned to two of 43 research readers for research reads. Features and categories were extracted from the clinical reports. This permitted the computation of interreader agreement between the research readers (R vs R) and between the research and clinical readers (R vs C). Twenty percent of images were also read twice by one of the research readers to permit the computation of intrareader agreement (R′ vs R′). DICOM = Digital Imaging and Communications in Medicine, LI-RADS = Liver Imaging Reporting and Data System, PHI = protected health information..

Statistical Analysis

Statisticians performed data analysis (T.W. and A.G., both with > 25 years of experience) using software (R; R Foundation for Statistical Computing). The research population and the reader questionnaire were summarized descriptively.

LI-RADS categories were combined into a four-category ordinal scale, with ascending risk of malignancy: LR-1 and LR-2; LR-3; LR-4; and LR-5, LR-M, and LR-TIV, where LR-M indicates probably malignant but not specific for HCC and LR-TIV indicates tumor in vein. It was necessary to pool categories with low frequency (LR-1 and LR-2) or that did not lend themselves to ordinal subranking (LR-5, LR-M, and LR-TIV) to allow for computation of overall agreement. Agreement was assessed using ICCs. Generally, ICCs less than 0.5, 0.5–0.75, 0.75–0.90, and greater than 0.90 indicate poor, moderate, good, and excellent agreement, respectively (27). Examination-level agreement was computed for research-versus-research reads and for research-versus-clinical reads. Subanalyses were performed for MRI and CT.

Binary agreement was computed using ICCs for LI-RADS categories dichotomized as probably or definitely malignant (LR-4, LR-5, LR-M, and LR-TIV vs LR-1, LR-2, and LR-3), LR-5, or LR-M; major features; and ancillary features present on at least 5% of images according to all reads (to ensure meaningful evaluation).

Nonparametric bootstrap analysis with per-case resampling was used to compute 95% CIs and to perform pairwise ICC comparisons (28,29). To our knowledge, there are no previously published data regarding clinical reads. Therefore, we considered the comparisons exploratory and did not correct for multiple statistical comparisons. $P < .05$ indicated statistical significance.

Results

CT and MRI examinations from 484 unique patients were included (Table 1). Seventy-four (15.3%) examinations originated

Table 1: Patient and Examination Characteristics

Characteristic	Result
Age (y)*	62 ± 10
Sex	
Male	328 (67.8)
Female	156 (32.2)
Modality	
CT	93 (19.2)
MRI with ECA	174 (36.0)
MRI with HBA	217 (44.8)
LI-RADS version 2018 categories based on features extracted from the clinical report	
LR-1	2 (0.4)
LR-2	35 (7.2)
LR-3	95 (19.6)
LR-4	153 (31.6)
LR-5	164 (33.9)
LR-M	27 (5.6)
LR-TIV	6 (1.2)
LR-NC	2 (0.4)
LI-RADS major features from the clinical report	
APHE	356 (73.6)
Washout	275 (56.8)
Enhancing capsule	117 (24.2)
Submitting radiologist	
C.W.H. (University of California San Diego; San Diego, Calif, United States)	118 (24.4)
V.C. (Montefiore Medical Center; New York, NY, United States)	211 (43.6)
S.L. (University of California Irvine; Orange, Calif, United States)	52 (10.7)
J.T.L. (University of Kentucky; Lexington, Ky, United States)	30 (6.2)
J.Y.C. (Yonsei University; Seoul, South Korea)	65 (13.4)
D.A. (Fundación Santa Fe de Bogotá; Bogotá, Colombia)	8 (1.7)

Note.—There were 484 examinations. Unless otherwise indicated, data are numbers of examinations; data in parentheses are percentages. APHE = arterial phase hyperenhancement, ECA = extracellular agent, HBA = hepatobiliary agent, LI-RADS = Liver Imaging Reporting and Data System, M = probably malignant, NC = not categorizable, TIV = tumor in vein.

* Mean age is ± SD.

from outside the United States. Patients included 156 (32.2%) women and 328 (67.8%) men, and they ranged in age from 21 to 95 years (mean age, 62 years ± 10 [SD]). Ninety-three (19.2%) CT and 391 (80.8%) MRI examinations were performed. MRI included 174 (36.0%) examinations with extracellular agents and 217 (44.8%) examinations with gadoxetic acid.

Research Reader Characteristics

The study included 43 research readers from 33 institutions and nine countries (Table 2), as follows: 33 readers were from the United States, two were from Canada, two were from Brazil, and one each was from China, South Korea, Colombia, India, Italy, and Australia. Forty-one readers (95%) reported fellowship training in abdominal imaging and the remaining two readers were current fellows in abdominal imaging. Thirty-nine readers (91%) self-identified themselves as experts in liver imaging. All readers mostly or almost exclusively read abdominal imaging in their daily clinical practice. Thirty-eight readers stated that their institution used LI-RADS in daily clinical practice (88%).

The readers reported an average of 11 years ± 6 of posttraining radiology experience. Thirty-five (81%) readers were in an academic setting, three (7%) readers were in private practice, and five (12%) readers were in a hybrid practice setting.

Each research reader interpreted 15–32 examinations (mean, 23 ± 3). Of those examinations, on average, 18 examinations were MRI and four examinations were CT. They reported spending 2–30 minutes per case (mean, 12 minutes ± 5).

Agreement for Modified Four-Category LI-RADS Scale

The agreement for the modified scale is summarized in Figure 2. Agreement was moderate for MRI (ICC, 0.68; 95% CI: 0.60, 0.74), CT (ICC, 0.68; 95% CI: 0.53, 0.80), and both modalities combined (ICC 0.68; 95%: CI 0.61, 0.73). For all modalities, better reader agreement was observed between research-versus-research reads than between research-versus-clinical reads (ICC, 0.68 [95% CI: 0.61, 0.73] vs 0.62 [95% CI: 0.56, 0.67], respectively; $P = .03$). Better reader agreement was also observed between research-versus-research reads for MRI (ICC, 0.68 [95% CI: 0.60, 0.74] vs 0.61 [95% CI: 0.54, 0.67], respectively; $P = .02$) but not for CT (ICC, 0.68 [95% CI: 0.53, 0.80] vs 0.66 [95% CI: 0.53, 0.76], respectively; $P = .66$).

Intrareader agreement for the modified scale was better than interreader agreement between research reads (ICC, 0.84 [95% CI: 0.74, 0.90] vs 0.68 [95% CI: 0.61, 0.73], respectively; $P = .002$).

Figure 3 and Figure S1 show examples of reader agreement and disagreement.

Agreement for Dichotomized LI-RADS Categories and for Individual Imaging Features

Agreement was moderate for dichotomized malignancy (LR-4, LR-5, LR-M, and LR-TIV: ICC, 0.63; 95% CI: 0.55, 0.70) and moderate for LR-5 versus any category but LR-5 (ICC, 0.58; 95% CI: 0.50, 0.66) (Fig 4). Agreement for LR-M versus any category but LR-M was poor (ICC, 0.46; 95% CI: 0.31, 0.61) (Fig 4). Better agreement for dichotomized malignancy was observed among research-versus-research reads compared with

Table 2: Research Reader Characteristics Based on Questionnaire Results

Question	Response Results
Country of primary affiliation	
U.S.	33
Canada	2
Brazil	2
China	1
South Korea	1
India	1
Columbia	1
Australia	1
Italy	1
Abdominal imaging fellowship	
Yes	41/43 (95)
No	2/43 (5)
Self-identified as expert in liver imaging	
Yes	39/43 (91)
No	4/43 (10)
Posttraining experience (y)	11 ± 6 (0–30)
Practice patterns	
Modalities used at institution	
Almost all MRI	8/43 (19)
More MRI than CT	20/43 (47)
Approximately equal use of MRI and CT	10/43 (23)
More CT than MRI	5/43 (12)
Almost all CT	0
MRI contrast agents used at institution	
Mostly extracellular agents	25/43 (58)
Approximately equal	5/43 (12)
Mostly gadoxetic acid	13/43 (30)
Liver cancer tumor board member	
Yes	36/43 (84)
No	7/43 (16)
Research read characteristics	
Self-reported time spent per case (min)	12 ± 5 (2–30)

Note.—For country of primary affiliation, the number of readers is shown. For all other categorical variables, data are numerators/denominators; data in parentheses are percentages. Mean data are ± SDs, with ranges in parentheses.

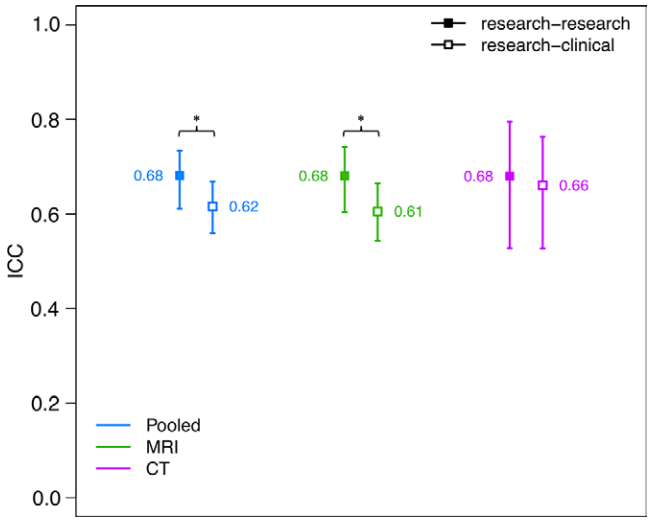

Figure 2: Plot shows intraclass correlation coefficient (ICC) reader agreement for modified four-category Liver Imaging Reporting and Data System (LI-RADS) version 2018 scale based on imaging modality. Agreement among research reads only (research-research; ■) and between research and clinical reads (research-clinical; □) are shown. Tails represent 95% CIs. * *P* value < .05 by nonparametric bootstrap with per-case resampling. Research-versus-research agreement pooled over both modalities and for MRI only was better than research-versus-clinical agreement.

research-versus-clinical reads (ICC, 0.63 [95% CI: 0.55, 0.70] vs 0.53 [95% CI: 0.46, 0.60], respectively; $P = .005$). No statistically significant differences in agreement among research reads compared with research-versus-clinical reads were observed for LR-5 (ICC, 0.58 [95% CI: 0.50, 0.66] vs 0.53 [95% CI: 0.47, 0.60], respectively; $P = .14$) or LR-M (ICC, 0.46 [95% CI: 0.31, 0.61] vs 0.46 [95% CI: 0.32, 0.61], respectively; $P = .94$).

Agreement was moderate for major features including arterial phase hyperenhancement (ICC, 0.65; 95% CI: 0.57, 0.72), washout (ICC, 0.53; 95% CI: 0.46, 0.60), and capsule (ICC, 0.50; 95% CI: 0.42, 0.58) (Fig 5). No difference in agreement was observed for research-versus-research reads compared with research-versus-clinical reads for arterial phase hyperenhancement (ICC, 0.65 [95% CI: 0.57, 0.72] vs 0.61 [95% CI: 0.54, 0.67], respectively; $P = .27$), washout (ICC, 0.53 [95% CI: 0.46, 0.60] vs 0.53 [95% CI: 0.46, 0.60], respectively; $P = .93$), or capsule (ICC, 0.50 [95% CI: 0.42, 0.58] vs 0.47 [95% CI: 0.38, 0.54], respectively; $P = .47$).

For ancillary features, agreement was moderate for restricted diffusion (ICC, 0.50; 95% CI: 0.42, 0.59) and for mild-moderate T2 hyperintensity (ICC, 0.58; 95% CI: 0.50, 0.66) (Fig 6). Agreement was poor for transitional phase hypointensity (ICC, 0.16; 95% CI: 0.03, 0.30) and hepatobiliary phase hypointensity (ICC, 0.44; 95% CI: 0.32, 0.55). Better agreement was observed for mild-moderate T2 hyperintensity among research reads than between research and clinical reads (ICC, 0.58 [95% CI: 0.50, 0.66] vs 0.46 [95% CI: 0.38, 0.54], respectively; $P = .01$). No differences in reader agreement were observed for the other ancillary features.

Discussion

Previous studies assessing the reader agreement of Liver Imaging Reporting and Data System (LI-RADS) have been limited by factors such as single-center nature, small number of readers, preselected images, and lack of comparison with clinical reads. We performed a large, multicenter, multireader study to begin to address knowledge gaps. The overall interreader agreement for a modified four-category LI-RADS scale was moderate among research reads (intraclass correlation coefficient [ICC], 0.68) and in comparing rescored clinical reads with research reads (ICC, 0.62). There was also moderate agreement for probably or definitely malignant categories (ICC, 0.63), for LR-5 (ICC, 0.58), and for all three major features (ICC, 0.50–0.65). For ancillary features, there was moderate agreement for restricted diffusion (ICC, 0.50) and mild-moderate T2 hyperintensity (ICC, 0.58), with poor agreement for

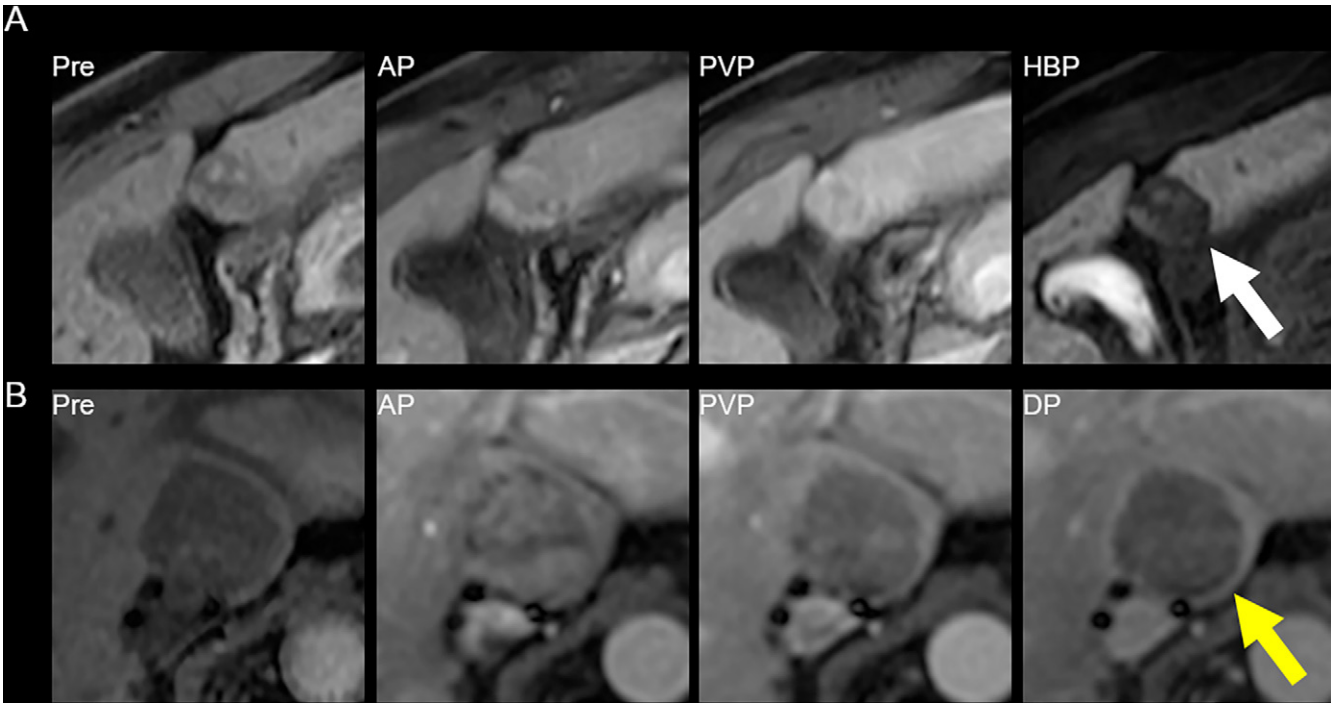

Figure 3: MRI scans show **(A)** reader disagreement and **(B)** reader agreement. **(A)** Gadoxetic acid–enhanced MRI scans in a 56-year-old male patient with cirrhosis secondary to hepatitis C. From left to right: contrast-unenhanced (Pre), arterial phase (AP), portal venous phase (PVP), and hepatobiliary phase (HBP) images. This 21-mm hepatobiliary phase hypointense observation (arrow) was characterized on the clinical read as having nonrim arterial phase hyperenhancement and washout appearance and was categorized as Liver Imaging Reporting and Data System (LI-RADS) category LR-5 (definitely hepatocellular carcinoma [HCC]). The first research reader characterized it as having a targetoid appearance and categorized it as LR-M (probably or definitely malignant, not specific for HCC). The second research reader characterized it as having no major features and paralleling the blood pool and categorized it as LR-2 (probably benign). It was subsequently resected and found to be a well-differentiated HCC. **(B)** Extracellular contrast–enhanced MRI scans in a 61-year-old female patient with cirrhosis secondary to hepatitis C. From left to right: contrast-unenhanced, arterial phase, portal venous phase, and delayed-phase (DP) images. This 31-mm observation (arrow) in the caudate lobe was characterized on the clinical read as having arterial phase hyperenhancement, washout appearance, and capsule appearance, and was categorized as LI-RADS category LR-5 (definitely HCC). Both research readers also categorized this observation as LR-5. The patient died of intracranial hemorrhage a few months later.

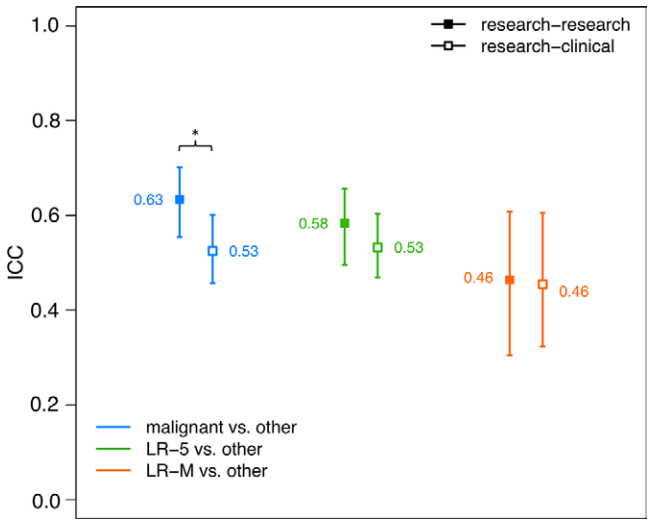

Figure 4: Plot shows intraclass correlation coefficient (ICC) reader agreement for dichotomized classification of Liver Imaging Reporting and Data System (LI-RADS) version 2018 for the following dichotomized categories: probably or definitely malignant versus other, LR-5 (definitely hepatocellular carcinoma [HCC]) versus other, and LR-M (probably or definitely malignant, not specific for HCC) versus other. Agreement among research reads only (research-research; ■) and between research and clinical reads (research-clinical; □) are shown. Tails represent 95% CIs. * $P < .05$ by nonparametric bootstrap with per-case resampling. Research-research agreement for malignant categories was better than research-clinical agreement.

transitional phase hypointensity (ICC, 0.16) and hepatobiliary phase hypointensity (ICC, 0.44).

A unique aspect of our study was the comparison between recomputed clinical reads and research reads. We found higher agreement between research reads than between research-versus-clinical reads for assignment of ordinal LI-RADS categories pooled over both modalities (ICC, 0.68 vs 0.62, respectively; $P = .03$) and for MRI (ICC, 0.68 vs 0.61, respectively; $P = .02$). Although this does not necessarily imply that agreement in the research environment will be higher than agreement in the clinical environment, these results indicate differences in interpretation between the clinical environment and the research environment that warrant further study. One possibility is that although the clinical reads were generally performed by subspecialty abdominal radiologists, many of the research readers were from LI-RADS committees and self-identified themselves as experts in liver imaging. In the clinical setting, prior imaging and reports may result in anchoring bias toward prior categorizations (30–32).

Several studies have provided important insights (25,26,33–35). Fowler et al (26) found an ICC of 0.67 for LI-RADS category assignment, which is similar to our result of 0.68. However, the study by Fowler et al reported agreement of 0.84–0.87 for the major features, which is higher than the agreement of 0.50–0.65 in our study. This might be due to of their use of selected

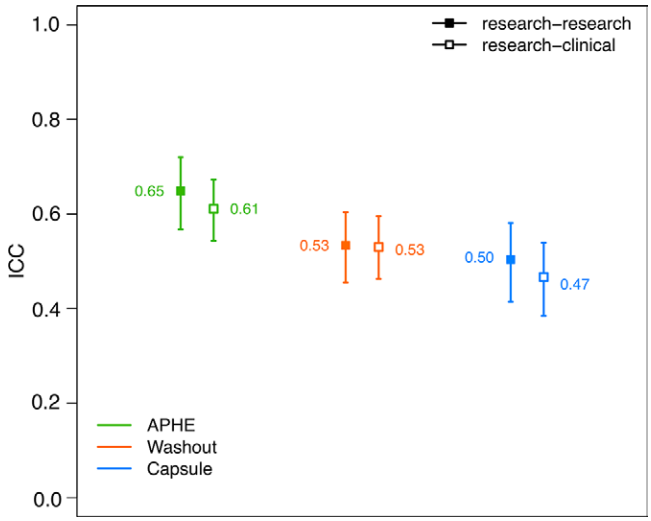

Figure 5: Plot shows intraclass correlation coefficient (ICC) reader agreement for Liver Imaging Reporting and Data System (LI-RADS) version 2018 arterial phase hyperenhancement (APHE), washout, and capsule for research reads only (research-research; ■) and between research and clinical reads (research-clinical; □). Tails represent 95% CIs. No differences in ICCs between research-versus-research reads compared with researcher-versus-clinical reads were observed.

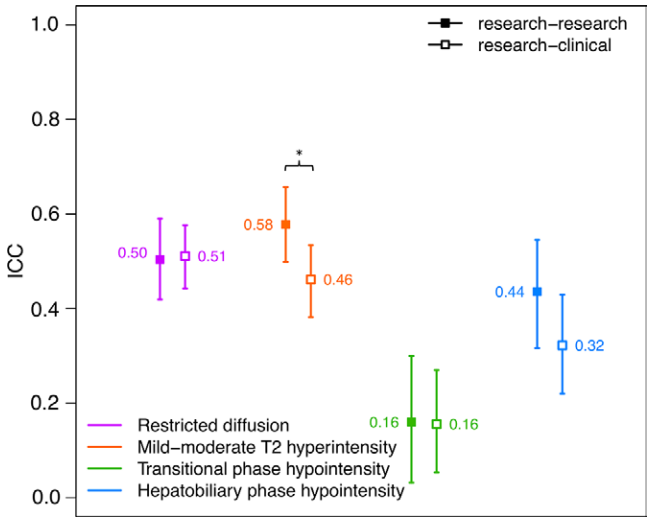

Figure 6: Plot shows intraclass correlation coefficient (ICC) reader agreement for Liver Imaging Reporting and Data System (LI-RADS) version 2018 ancillary features with sufficient frequency for analysis, which included restricted diffusion, mild-moderate T2 hyperintensity, transitional phase hypointensity, and hepatobiliary phase hypointensity. These four features are only visible at MRI. Agreement among research reads only (research-research; ■) and between research and clinical reads (research-clinical; □) are shown. Tails represent 95% CIs. * $P < .05$ by nonparametric bootstrap with per-case resampling. Research-versus-research agreement for mild-moderate T2 hyperintensity was better than research-versus-clinical agreement.

image sets versus our use of scrollable examinations, which may have shown the imaging features more clearly. Kang et al (25) performed a meta-analysis of 15 studies and found a pooled κ of 0.66–0.72 for the major features, compared with the ICC of 0.50–0.65 in our study. These variations may be related to differences in study design; Kang et al reported substantial study heterogeneity within the included studies. Similar to the study by Kang et al and other previous studies, our study found that nonrim arterial phase hyperenhancement had the highest reader agreement of the major features.

Our study had several limitations. First, although we did have international participation in this study, most of our examinations and readers were from academic medical centers in North America. Additionally, 91% of our research readers were self-reported experts in liver imaging, and only 7% were in private practice. Thus, further evaluation of LI-RADS among community radiologists and medical centers outside of North America should be the focus of future work. Our study did not assess agreement of treatment response categories, and therefore our results only generalized to untreated observations. Annotating the observation may have introduced bias based on the selected image. We could only assess the interpretations that were recomputed using LI-RADS version 2018, excluding features that depended on prior comparisons rather than the clinically reported categories. The reader agreement for subthreshold and threshold growth could not be assessed. In addition, the number of possible pairs of research readers exceeded the number of examinations, which precluded meaningful evaluation of the effect of reader characteristics on agreement. Finally, we could not directly evaluate agreement between clinical reads. It is possible that clinical agreement is similar to research agreement, just that clinical reads are different from research reads.

In conclusion, Liver Imaging Reporting and Data System (LI-RADS) version 2018 generally has moderate agreement for observation categorization and feature characterization. Future research is needed to identify methods for reducing variability among readers, such as training, structured reporting, automated category computation based on reported features, or development of computer-aided categorization. In the meantime, it is important to be mindful of this variability because it can substantially impact patient care, and selected patients should be referred to multidisciplinary tumor boards when feasible for consensus diagnostic and treatment decisions. At institutions without multidisciplinary tumor boards, double reading and/or referral of these patients to centers with such tumor boards should be considered. There are differences in interpretation between the research and clinical environments that warrant further study. Future research studies should also focus on the diagnostic performance of LI-RADS in the clinical setting, especially among community radiologists and in medical centers outside of North America, which are important knowledge gaps in the validation of LI-RADS.

Acknowledgments: We gratefully acknowledge the contributions of the late Dr Christopher Kagay to the study design, data acquisition, and manuscript revisions.

Author contributions: Guarantors of integrity of entire study, **C.W.H., C.P., J.T.L., K.G., B.B., D.T.M.**; 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; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, **C.W.H., C.P., J.B., D.A.A., M.L.D., K.M.E., V.R.S., S.B., K.J.F., C.B.S.**; clinical studies, **C.W.H., V.C., J.Y.C., S.L., C.P., C.L., J.T.L., J.W.O., D.A.A., M.M.L., M.S.D., W.M., A.R., S.C.L., E.M.H., M.R.B., M.A.O., A.T., A.F., E.A.C., M.T.C., K.M.E., V.R.S., R.M.M., N.H., S.B., K.G., A.K., J.W., I.C., B.B., G.M.C., D.T.M., K.J.F.**; experimental

studies, **T.D., A.F., V.R.S., K.B., R.M.M., N.H., J.W., K.J.F.**; statistical analysis, **C.W.H., T.W., A.G., V.R.S., J.W., B.B., K.J.F.**; and manuscript editing, **C.W.H., V.C., S.L., T.D., T.W., A.G., J.B., R.K., C.L., J.T.L., J.W.O., D.A.A., M.M.L., M.S.D., A.R., S.C.L., A.S.K., E.M.H., M.R.B., G.B., M.L.D., M.A.O., A.T., M.C., A.F., E.A.C., M.T.C., J.P.M., B.K., K.M.E., V.R.S., K.B., R.M.M., N.H., S.B., R.A., K.G., C.R.K., A.K., J.W., I.C., B.B., M.G., G.M.C., D.T.M., K.J.F., C.B.S.**

Disclosures of conflicts of interest: **C.W.H.** No relevant relationships. **V.C.** Consulting fees from Bayer. **J.Y.C.** No relevant relationships. **S.L.** No relevant relationships. **C.P.** No relevant relationships. **T.D.** No relevant relationships. **T.W.** No relevant relationships. **A.G.** No relevant relationships. **J.B.** No relevant relationships. **R.K.** No relevant relationships. **C.L.** No relevant relationships. **J.T.L.** No relevant relationships. **J.W.O.** LI-RADS Rad-Path Workgroup Co-chair. **D.A.A.** No relevant relationships. **M.M.L.** No relevant relationships. **M.S.D.** Royalties from Wolters-Kluwer, UpToDate.com; treasurer on Board of Directors for Society of Advanced Body Imaging; member of *Radiology* editorial board. **W.M.** No relevant relationships. **A.R.** No relevant relationships. **S.C.L.** No relevant relationships. **A.S.K.** No relevant relationships. **E.M.H.** No relevant relationships. **M.R.B.** Grant funding to author institution from NCI, Siemens Healthineers, Madrigal Pharmaceuticals, NGM Biopharmaceuticals, Carmot Therapeutics, Corcept Therapeutics.; member of *Radiology* editorial board **G.B.** Consulting fees from Bayer; payment for lectures from Bayer, Guerbet, Bracco, AstraZeneca; support for meeting attendance from Bayer, AstraZeneca. **M.L.D.** No relevant relationships. **M.A.O.** No relevant relationships. **A.T.** Grants from Institut de Valorisation des Donnees, Canadian Institutes of Health Research; consulting fees from Onco-Tech; LI-RADS steering committee member; LI-RADS International Working Group member; equipment loan from Siemens Healthcare for Onco-Tech and CIHR grants. **M.C.** No relevant relationships. **A.F.** LI-RADS Technique Working Group co-chair. **E.A.C.** No relevant relationships. **M.T.C.** No relevant relationships. **J.P.M.** No relevant relationships. **B.K.** No relevant relationships. **K.M.E.** No relevant relationships. **V.R.S.** No relevant relationships. **K.B.** No relevant relationships. **R.M.M.** No relevant relationships. **N.H.** Consulting fees from Guerbet; payment for lecture from Bayer. **S.B.** No relevant relationships. **R.A.** No relevant relationships. **K.G.** No relevant relationships. **C.R.K.** No relevant relationships. **A.K.** Research grants from GE Healthcare, Philips Healthcare, PanCAN, Bayer; advisory board membership, Bayer; honorarium from Texas Radiological Society. **J.W.** No relevant relationships. **I.C.** No relevant relationships. **B.B.** No relevant relationships. **M.G.** No relevant relationships. **G.M.C.** 2021 Bayer Healthcare/RSNA Research Fellow grant recipient. **D.T.M.** No relevant relationships. **K.J.F.** Grants from Bayer, Pfizer, Median; consulting fees from Epigenomics, Bayer; payment for lectures from CME Talks; payment for expert testimony; participant on DataSafety or Advisory Board, Bayer; member of the RSNA Editorial Board, ACR panel chair, ACR LI-RADS, SAR portfolio director; unpaid board member for Quantix Bio. **C.B.S.** Research grants from ACR, Bayer, Foundation of the NIH, GE, Gilead, Pfizer, Philips, Siemens; lab service agreements from Enanta, Gilead, ICON, Intercept, Nusirt, Shire, Synageva, Takeda; royalties from Medscape, Wolters-Kluwer; institutional consulting representative for AMRA, BMS, Exact Sciences, IBM-Watson, Pfizer; personal consulting for Blade, Boehringer, Epigenomics, Guerbet; honoraria for educational symposia from Japanese Society of Radiology, Stanford, M.D. Anderson; advisory board member for Quantix Bio; Chief Medical Officer for Livivos; stock options in Livivos; member of LI-RADS Steering Committee.

References

1. Bruix J, Sherman M; American Association for the Study of Liver Diseases. Management of hepatocellular carcinoma: an update. Hepatology 2011;53(3):1020–1022.
2. Elsayes KM, Kielar AZ, Agrons MM, et al. Liver Imaging Reporting and Data System: an expert consensus statement. J Hepatocell Carcinoma 2017;4:29–39.
3. Elsayes KM, Kielar AZ, Chernyak V, et al. LI-RADS: a conceptual and historical review from its beginning to its recent integration into AASLD clinical practice guidance. J Hepatocell Carcinoma 2019;6:49–69.
4. Chen N, Motosugi U, Morisaka H, et al. Added Value of a Gadoxetic Acid-enhanced Hepatocyte-phase Image to the LI-RADS System for Diagnosing Hepatocellular Carcinoma. Magn Reson Med Sci 2016;15(1):49–59.
5. Choi SH, Byun JH, Kim SY, et al. Liver Imaging Reporting and Data System v2014 With Gadoxetate Disodium-Enhanced Magnetic Resonance Imaging: Validation of LI-RADS Category 4 and 5 Criteria. Invest Radiol 2016;51(8):483–490.
6. Cha DI, Jang KM, Kim SH, Kang TW, Song KD. Liver Imaging Reporting and Data System on CT and gadoxetic acid-enhanced MRI with diffusion-weighted imaging. Eur Radiol 2017;27(10):4394–4405.
7. Abd Alkhalik Basha M, Abd El Aziz El Sammak D, El Sammak AA. Diagnostic efficacy of the Liver Imaging-Reporting and Data System (LI-RADS) with CT imaging in categorising small nodules (10-20 mm) detected in the cirrhotic liver at screening ultrasound. Clin Radiol 2017;72(10):901.e1–901.e11.
8. Kim YY, An C, Kim S, Kim MJ. Diagnostic accuracy of prospective application of the Liver Imaging Reporting and Data System (LI-RADS) in gadoxetate-enhanced MRI. Eur Radiol 2018;28(5):2038–2046.
9. Liu W, Qin J, Guo R, et al. Accuracy of the diagnostic evaluation of hepatocellular carcinoma with LI-RADS. Acta Radiol 2018;59(2):140–146.
10. van der Pol CB, Lim CS, Sirlin CB, et al. Accuracy of the Liver Imaging Reporting and Data System in Computed Tomography and Magnetic Resonance Image Analysis of Hepatocellular Carcinoma or Overall Malignancy-A Systematic Review. Gastroenterology 2019;156(4):976–986.
11. Tanabe M, Kanki A, Wolfson T, et al. Imaging Outcomes of Liver Imaging Reporting and Data System Version 2014 Category 2, 3, and 4 Observations Detected at CT and MR Imaging. Radiology 2016;281(1):129–139.
12. Hong CW, Park CC, Mamidipalli A, et al. Longitudinal evolution of CT and MRI LI-RADS v2014 category 1, 2, 3, and 4 observations. Eur Radiol 2019;29(9):5073–5081.
13. Cerny M, Bergeron C, Billiard JS, et al. LI-RADS for MR Imaging Diagnosis of Hepatocellular Carcinoma: Performance of Major and Ancillary Features. Radiology 2018;288(1):118–128.
14. Kim YY, Kim MJ, Kim EH, Roh YH, An C. Hepatocellular Carcinoma versus Other Hepatic Malignancy in Cirrhosis: Performance of LI-RADS Version 2018. Radiology 2019;291(1):72–80.
15. Lee S, Kim SS, Roh YH, Choi JY, Park MS, Kim MJ. Diagnostic Performance of CT/MRI Liver Imaging Reporting and Data System v2017 for Hepatocellular Carcinoma: A Systematic Review and Meta-Analysis. Liver Int 2020;40(6):1488–1497.
16. Lee SM, Lee JM, Ahn SJ, Kang HJ, Yang HK, Yoon JH. LI-RADS Version 2017 versus Version 2018: Diagnosis of Hepatocellular Carcinoma on Gadoxetate Disodium-enhanced MRI. Radiology 2019;292(3):655–663.
17. Chen J, Kuang S, Zhang Y, et al. Increasing the sensitivity of LI-RADS v2018 for diagnosis of small (10-19 mm) HCC on extracellular contrast-enhanced MRI. Abdom Radiol (NY) 2021;46(4):1530–1542.
18. Zhang Y, Tang W, Xie S, et al. The role of lesion hypointensity on gadobenate dimeglumine-enhanced hepatobiliary phase MRI as an additional major imaging feature for HCC classification using LI-RADS v2018 criteria. Eur Radiol 2021;31(10):7715–7724.
19. Cha DI, Choi GS, Kim YK, et al. Extracellular contrast-enhanced MRI with diffusion-weighted imaging for HCC diagnosis: prospective comparison with gadoxetic acid using LI-RADS. Eur Radiol 2020;30(7):3723–3734.
20. Lee CM, Choi SH, Byun JH, et al. Combined computed tomography and magnetic resonance imaging improves diagnosis of hepatocellular carcinoma ≤ 3.0 cm. Hepatol Int 2021;15(3):676–684.
21. Chung JW, Yu JS, Choi JM, Cho ES, Kim JH, Chung JJ. Subtraction Images From Portal Venous Phase Gadoxetic Acid-Enhanced MRI for Observing Washout and Enhancing Capsule Features in LI-RADS Version 2018. AJR Am J Roentgenol 2020;214(1):72–80.
22. Hwang SH, Park S, Han K, Choi JY, Park YN, Park MS. Optimal lexicon of gadoxetic acid-enhanced magnetic resonance imaging for the diagnosis of hepa-tocellular carcinoma modified from LI-RADS. Abdom Radiol (NY) 2019;44(9):3078–3088.
23. Min JH, Kim JM, Kim YK, et al. A modified LI-RADS: targetoid tumors with enhancing capsule can be diagnosed as HCC instead of LR-M lesions. Eur Radiol 2022;32(2):912–922.
24. Chen J, Zhou J, Kuang S, et al. Liver Imaging Reporting and Data System Category 5: MRI Predictors of Microvascular Invasion and Recurrence After Hepatectomy for Hepatocellular Carcinoma. AJR Am J Roentgenol 2019;213(4):821–830.
25. Kang JH, Choi SH, Lee JS, et al. Interreader Agreement of Liver Imaging Reporting and Data System on MRI: A Systematic Review and Meta-Analysis. J Magn Reson Imaging 2020;52(3):795–804.
26. Fowler KJ, Tang A, Santillan C, et al. Interreader Reliability of LI-RADS Version 2014 Algorithm and Imaging Features for Diagnosis of Hepatocellular Carcinoma: A Large International Multireader Study. Radiology 2018;286(1):173–185.
27. Koo TK, Li MY. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 2016;15(2):155–163. [Published correction appears in J Chiropr Med 2017;16(4):346.]
28. Streiner DL, Norman GR. Health Measurement Scales: A Practical Guide to Their Development and Use. 2nd ed. New York, NY: Oxford University Press, 1994.
29. Hinkley DV. Bootstrap Methods. J R Stat Soc Ser B Methodol 1988;50(3):321–337.

30. Lee CS, Nagy PG, Weaver SJ, Newman-Toker DE. Cognitive and system factors contributing to diagnostic errors in radiology. AJR Am J Roentgenol 2013;201(3):611–617.
31. Bruno MA, Walker EA, Abujudeh HH. Understanding and Confronting Our Mistakes: The Epidemiology of Error in Radiology and Strategies for Error Reduction. RadioGraphics 2015;35(6):1668–1676.
32. Busby LP, Courtier JL, Glastonbury CM. Bias in Radiology: The How and Why of Misses and Misinterpretations. RadioGraphics 2018;38(1):236–247.
33. Abdel Razek AAK, El-Serougy LG, Saleh GA, Abd El-Wahab R, Shabana W. Interobserver Agreement of Magnetic Resonance Imaging of Liver Imaging Reporting and Data System Version 2018. J Comput Assist Tomogr 2020;44(1):118–123.
34. Schellhaas B, Hammon M, Strobel D, et al. Interobserver and intermodality agreement of standardized algorithms for non-invasive diagnosis of hepatocellular carcinoma in high-risk patients: CEUS-LI-RADS versus MRI-LI-RADS. Eur Radiol 2018;28(10):4254–4264.
35. Lim K, Kwon H, Cho J. Inter-reader agreement and imaging-pathology correlation of the LI-RADS M on gadoxetic acid-enhanced magnetic resonance imaging: efforts to improve diagnostic performance. Abdom Radiol (NY) 2020;45(8):2430–2439.