European Radiology (2021) 31:6856–6867
https://doi.org/10.1007/s00330-021-07815-y

GASTROINTESTINAL

Inter-reader reliability of CT Liver Imaging Reporting and Data System according to imaging analysis methodology: a systematic review and meta-analysis

Ji Hun Kang[1] · Sang Hyun Choi[2] · Ji Sung Lee[3,4] · Kyung Won Kim[2] · So Yeon Kim[2] · Seung Soo Lee[2] · Jae Ho Byun[2]

Received: 13 November 2020 / Revised: 1 January 2021 / Accepted: 18 February 2021 / Published online: 13 March 2021

Abstract

Objectives To establish inter-reader reliability of CT Liver Imaging Reporting and Data System (LI-RADS) and explore factors that affect it.

Methods MEDLINE and EMBASE databases were searched from January 2014 to March 2020 to identify original articles reporting the inter-reader reliability of CT LI-RADS. The imaging analysis methodology of each study was identified, and pooled intraclass correlation coefficient (ICC) or kappa values (κ) were calculated for lesion size, major features (arterial-phase hyperenhancement [APHE], nonperipheral washout [WO], and enhancing capsule [EC]), and LI-RADS categorization (LR) using random-effects models. Subgroup analyses of pooled κ were performed for the number of readers, average reader experience, differences in reader experience, and LI-RADS version.

Results In the 12 included studies, the pooled ICC or κ of lesion size, APHE, WO, EC, and LR were 0.99 (0.96−1.00), 0.69 (0.58–0.81), 0.67 (0.53–0.82), 0.65 (0.54–0.76), and 0.70 (0.59–0.82), respectively. The experience and number of readers varied: studies using readers with ≥ 10 years of experience showed significantly higher κ for LR (0.82 vs. 0.45, $p = 0.01$) than those with < 10 years of reader experience. Studies with multiple readers including inexperienced readers showed significantly lower κ for APHE (0.55 vs. 0.76, $p = 0.04$) and LR (0.45 vs. 0.79, $p = 0.02$) than those with all experienced readers.

Conclusions CT LI-RADS showed substantial inter-reader reliability for major features and LR. Inter-reader reliability differed significantly according to average reader experience and differences in reader experience. Reported results for inter-reader reliability of CT LI-RADS should be understood with consideration of the imaging analysis methodology.

Key Points

- *The CT Liver Imaging Reporting and Data System (LI-RADS) provides substantial inter-reader reliability for three major features and category assignment.*
- *The imaging analysis methodology varied across studies.*
- *The inter-reader reliability of CT LI-RADS differed significantly according to the average reader experience and the difference in reader experience.*

Keywords Liver · Hepatocellular carcinoma · Multidetector computed tomography · Reproducibility of results · Meta-analysis

✉ Sang Hyun Choi
edwardchoi83@gmail.com

[1] Department of Radiology, Hanyang University College of Medicine, Hanyang University Guri Hospital, Guri-si, Gyeonggi-do, Republic of Korea

[2] Department of Radiology and Research Institute of Radiology, University of Ulsan College of Medicine, Asan Medical Center, 88 Olympic-Ro 43-Gil, Songpa-Gu, Seoul 05505, Republic of Korea

[3] Department of Clinical Epidemiology and Biostatistics, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

[4] Clinical Research Center, Asan Institute for Life Sciences, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

Abbreviations

APHE	Arterial-phase hyperenhancement
CI	Confidence interval
EC	Enhancing capsule
HCC	Hepatocellular carcinoma
ICC	Intraclass correlation coefficient
LI-RADS	Liver Imaging Reporting and Data System
WO	Nonperipheral washout

Introduction

The Liver Imaging Reporting and Data System (LI-RADS) was introduced in 2011 [1] and recently updated in 2018 to standardize the performance of liver imaging in patients at risk for hepatocellular carcinoma (HCC) [2, 3]. LI-RADS provides a standardized lexicon for imaging features of HCC, as well as criteria for ordinal categories (i.e., LR-1 to LR-5, according to the likelihood of benignity or HCC). Unlike other malignant tumors, HCC in patients at risk can be diagnosed noninvasively on the basis of imaging features on dynamic CT or MRI, without mandatory pathologic confirmation [4]. It is therefore very important to standardize the imaging diagnosis of HCC [5].

Given the wide adoption of LI-RADS in research and clinical practice, extensive evaluation of the diagnostic accuracy of LI-RADS has been reported [6–8], but variable imaging protocols and the lack of using standardized lexicon may challenge the synthesis of solid evidence [9]. A recent multi-center, multi-reader study found that CT LI-RADS demonstrated substantial to almost perfect inter-reader reliability, with intraclass correlation coefficients (ICCs) of 0.67 for LI-RADS category assignment and 0.79 to 0.86 for major features [9]. In addition, this previous study showed that the inter-reader reliability of LI-RADS was not significantly affected by LI-RADS familiarity or years of experience [9]. Although this multi-center, multi-reader study determined the inter-reader reliability for CT LI-RADS, variable results for inter-reader reliability of CT LI-RADS have been reported after this study [10–13], with kappa values (κ) for LI-RADS category assignment ranging from 0.44 to 0.90.

Recently, a meta-analysis of inter-reader reliability of MRI LI-RADS has been published, reporting overall substantial agreement [14]. However, that of CT LI-RADS has not yet been done. Considering the fact that discrepancy rates in imaging tests can be influenced by various imaging analysis factors, as well as reader characteristics [15, 16], we hypothesized that differences in imaging analysis methodology between each study might be partly responsible for these heterogeneous results for the inter-reader reliability of CT LI-RADS. Therefore, we aimed to establish inter-reader reliability of CT LI-RADS and explore factors that affect it.

Materials and methods

This study was conducted and reported according to the guidelines for Meta-analysis of Observational Studies in Epidemiology [17] and Preferred Reporting Items for Systematic Reviews and Meta-Analyses [18, 19].

Literature search

A systematic search was performed in MEDLINE and EMBASE databases to find original studies reporting the inter-reader reliability of LI-RADS categorizations and major imaging features of LI-RADS for the diagnosis of HCC using CT. The search query was developed to provide a sensitive search of potentially eligible articles (Supplementary Table 1). The search was performed for articles published from January 2014, to include original studies using LI-RADS v2014 or later versions of LI-RADS. The literature search was updated until March 2020. The search was limited to studies published in English and those involving human subjects. The bibliographies of the included articles were also screened to expand the scope of the search and to prevent other potentially relevant studies from being omitted.

Eligibility criteria

Studies were included when all of the following criteria were met: (a) population—patients at risk for HCC with a focal observation, i.e., adult patients with cirrhosis or chronic viral hepatitis [20]; (b) index test—dynamic contrast-enhanced CT; (c) comparator—no requirements; (d) outcome—inter-reader reliability of major imaging features and LI-RADS categorization; and (e) study design—any type of study including observational studies and clinical trials. Studies were excluded when any of the following criteria were met: (a) case reports, meta-analyses, review articles, letters, comments, and conference abstracts; (b) studies with overlapping patient cohorts and data; (c) studies not related to the field of interest of this study; and (d) studies with insufficient data to determine inter-reader reliability. Studies were independently screened by two reviewers using their titles and abstracts according to the inclusion and exclusion criteria. After exclusion of ineligible studies, we performed a full-text review of the remaining potentially eligible studies. If any disagreement was present between the two reviewers, the studies were re-evaluated at a consensus meeting.

Data extraction

Using predefined data forms, the following data were extracted from the included studies: (a) study characteristics—author, year of publication, publishing country, study design, study type, and subject enrollment method; (b) demographic and clinical characteristics—number of patients, patients' age, number, and type of lesions, and type of reference standard; (c) CT techniques—number of detectors, multiphase sequence, and slice thickness; (d) methodology of imaging analysis—number of readers, experience of each reader, difference in reader experience, independent review, clarity of blinding to reference standard in the review, and version of LI-RADS used; and (e) study outcomes—inter-reader reliability according to lesion size, major features (arterial-phase hyperenhancement [APHE], nonperipheral washout [WO], and enhancing capsule [EC]), and LI-RADS categorization. To determine inter-reader reliability, the ICCs for continuous variables and κ for categorical variables with standard errors were extracted for each major feature and LI-RADS categorization. Data extraction was performed independently by the two reviewers, and any discrepancies between them were resolved at a consensus meeting.

Quality assessment

Study quality was assessed using the Guidelines for Reporting Reliability and Agreement studies [21]. Important elements included the following seven domains: descriptions of the index test (CT techniques), study subjects (recruitment methods and demographic characteristics), readers (number and experience level), reading process (availability of clinical information and independent review), clarity of the blinding review, statistical analysis, and actual number of subjects and observations. Each category was scored as high quality if it was described in sufficient detail in the article with no potential bias. The study quality was assessed independently by the two reviewers, with any discrepancies between them being resolved at a consensus meeting.

Data synthesis and statistical analysis

To calculate meta-analytic summary estimates, the ICC with standard error was summarized for lesion size, and κ with standard error was summarized for APHE, WO, EC, and LI-RADS categorization from each study. If the original study did not report standard error, it was estimated from the 95% confidence interval (CI). The meta-analytic pooled ICC and κ with 95% CI were calculated using the DerSimonian-Laird random-effects model with or without Knapp and Hartung adjustment [22]. ICC and κ were categorized according to Landis and Koch as follows: < 0.20, poor; 0.21–0.40, fair; 0.41–0.60, moderate; 0.61–0.80, substantial; and 0.81–1.00, almost perfect reliability [23]. Heterogeneity was assessed using the Cochran Q-test and I^2 statistics [24], with $I^2 > 50\%$ or $p < 0.10$ in the Cochran Q-test being considered to indicate substantial heterogeneity.

To evaluate the inter-reader reliability of LI-RADS according to the imaging analysis methodology, we performed subgroup analyses according to the following variables: (a) number of readers (two readers vs. more than two readers); (b) average reader experience (≥ 10 years of experience in abdominal/liver imaging vs. < 10 years of experience); and (c) difference in reader experience (all experienced readers vs. multiple readers with inexperienced readers, i.e., < 5 years of post-fellowship experience). In addition, we performed subgroup analysis according to the version of LI-RADS used (v2014, v2017, or v2018).

Meta-regression analysis was performed to explore the causes of study heterogeneity, which included the following covariates: study design (prospective vs. retrospective), study type (cohort vs. case-control), subject enrollment (consecutive vs. selective), number of CT detectors (≥ 64 channels in all included CT vs. others), dynamic CT sequence (quad-phase including unenhanced, arterial phase, portal venous phase, and delayed or equilibrium phase vs. triple-phase), CT slice thickness (≤ 3 mm vs. others), and clarity of blinding to reference standard during the review (clear vs. unclear).

Funnel plots and rank tests were used to assess the presence of any publication bias. R version 3.6.3 with the "metafor" package was used to perform the analyses.

Results

Literature search

The systematic literature search initially identified 263 articles (Fig. 1). After removing 109 duplicate articles, 154 articles were screened by their titles and abstracts, and 99 articles were excluded. Full-text reviews were performed for 55 potentially eligible articles, and 43 articles were excluded. Finally, 12 original articles with a total of 2285 lesions were included in this study [9–13, 25–31].

Characteristics of the included studies

The detailed characteristics of the included studies are summarized in Table 1. Eleven studies were retrospective [9–13, 26–31], and one study was prospective [25]. Eleven studies were cohort studies [9–13, 25–28, 30, 31], and one was a case-control study [29]. All of the included studies were analyzed independently by the readers. Six studies used histopathology for the reference standard [11–13, 26, 29, 30], four used a combination of histopathology and clinical follow-up [10, 25, 27, 31], and two did not use a reference standard because

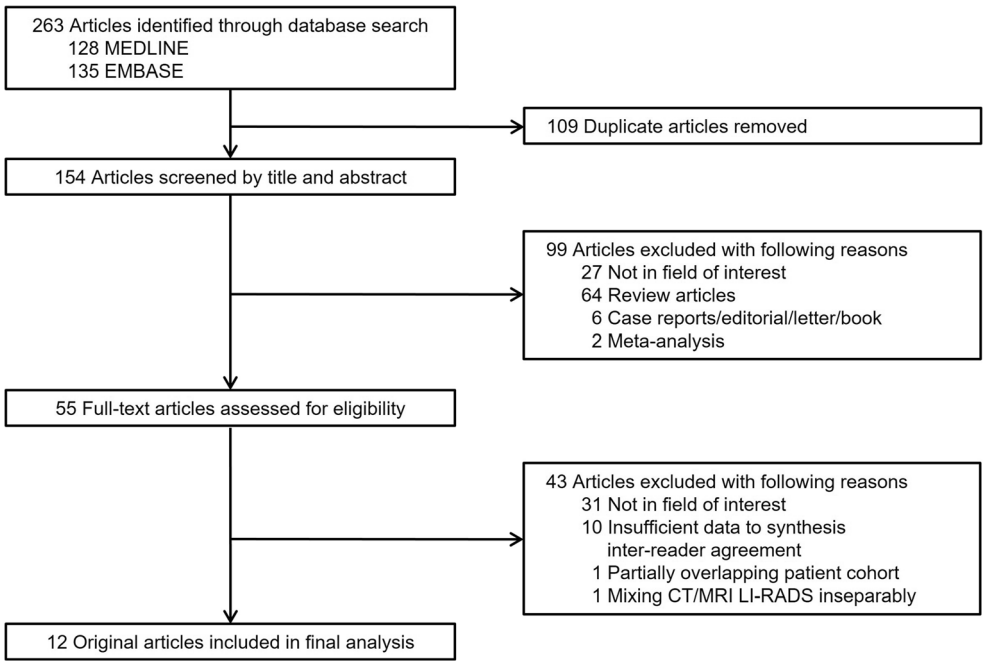

Fig. 1 Flow diagram of the study selection process. LI-RADS, Liver Imaging Reporting and Data System

they focused on inter-reader reliability rather than diagnostic accuracy [9, 28]. The readers in 10 studies were blinded to the reference standard [10–13, 25–27, 29–31], whereas the other two studies were unclear on blinding [9, 28].

Imaging analysis methodologies of the included studies

The imaging analysis methodologies of each study are summarized in Table 2. Nine studies used two readers [10–12, 25, 27–31], and three studies used more than two [9, 13, 26]. The experience level of the readers was variable, ranging from trainees to 23 years of experience in liver imaging. Of the 12 included studies, eight stated the experience of each reader [10–12, 25, 26, 28, 30, 31], with the average reader experience being 10.8 years in abdominal/liver imaging. Three studies included both experienced readers and inexperienced readers [12, 13, 31]. Nine studies used LI-RADS v2014 [9, 13, 25–31], two studies used v2017 [10, 31], and one study used v2018 [12].

Study quality

All included studies showed a quality score of five or more for the seven criteria evaluated. Description of the index test was lacking in two studies [9, 29], and readers' experience levels were not reported in two studies [27, 29]. In addition, the presence of blinding during the review was unclear in two studies [9, 28]. Further details on the study quality are provided in Supplementary Figure 1.

Meta-analytic pooled inter-reader reliability of LI-RADS

The meta-analytic pooled estimates of inter-reader reliability of LI-RADS are summarized in Fig. 2. For lesion size, the ICCs ranged from 0.74 to 0.99, and the meta-analytic pooled ICC was 0.99 (95% CI, 0.96−1.00), which was close to perfect reliability. Among the three major features, the highest meta-analytic pooled κ was shown by APHE (0.69; 95% CI, 0.58–0.81), followed by WO (0.67; 95% CI, 0.53–0.82). All three major features showed substantial inter-reader reliability. The κ for the LI-RADS categorization ranged from 0.44 to 0.90, with a meta-analytic pooled κ of 0.70 (95% CI, 0.59–0.82), showing substantial inter-reader reliability.

Subgroup analysis according to the imaging analysis methodology

Subgroup analyses of inter-reader reliability for the major features and LI-RADS categorization according to the imaging analysis methodology are summarized in Table 3. Both studies with two readers and those with more than two readers showed moderate to substantial inter-reader reliability for the three major features (κ = 0.64–0.71) and LI-RADS categorization (κ = 0.56–0.64). Regarding the reader experience, the meta-analytic pooled κ for LI-RADS categorization was significantly higher in studies using readers with ≥ 10 years of experience than in those using readers with < 10 years of experience (p = 0.01). Other meta-analytic pooled estimates, including lesion

Table 1 Characteristics of the included studies

Author (year of publication)	Study design	Study type	Subject enrollment	Country	No. of patients	Mean age (range)	No. of lesions	Lesions types (no.)	CT technique	Reference standard	Blinding to reference standard*	Analyzed features
Alhasan A et al (2019)	Retrospective	Cohort	Consecutive	Canada	59	63.2 (NA)	104	HCC (72), NHM (3), benign (29)	MDCT (16, 64 channel) with quad-phase CE, 2.5–3.0 mm thickness	Pathology or CCRS	Yes	APHE, WO, EC, LR
Basha MAA et al (2018)	Prospective	Cohort	Consecutive	Egypt	240	61.5 (35–79)	247	NA	MDCT (64, 128 channel) with quad-phase CE, 5.0 mm thickness	Pathology or CCRS	Yes	APHE, WO, EC, LR
Cha DI et al (2017)	Retrospective	Cohort	Selective	Korea	421	57.0 (20–82)	445	HCC (397), NHM (31), benign (17)	MDCT (64 channel) with quad-phase CE, 5.0 mm thickness	Pathology	Yes	APHE, WO, EC, LR
Chen N et al (2016)	Retrospective	Cohort	Consecutive	Japan	139	68.0 (18–89)	139	HCC (111), benign (28)	MDCT (64 channel) with quad-phase CE, 5.0 mm thickness	Pathology or CCRS	Yes	LR
Chernyak V et al (2018)	Retrospective	Cohort	Consecutive	America	42	62.2 (NA)	50	NA	MDCT (16, 64 channel) with triple or quad-phase CE, 2.0–3.0 mm thickness	NA	Unclear	Size, APHE, WO, EC, LR
Ehman EC et al (2016)	Retrospective	Case control	Selective	America	134	58.2 (NA)	141	HCC (141)	Quad-phase CE	Pathology	Yes	Size
Fowler KJ et al (2018)	Retrospective	Cohort	Selective	America	NA	NA	382	NA	NA	NA	Unclear	APHE, WO, EC, LR
Joo I et al (2017)	Retrospective	Cohort	Selective	Korea	186	57.9 (29–85)	216	HCC (216)	MDCT (8, 16, 64 channel) with quad-phase CE, 2.5–3.0 mm thickness	Pathology	Yes	Size, APHE, WO, EC
Kim SS et al (2019)	Retrospective	Cohort	Selective	Korea	106	59.5 (38–79)	112	HCC (112)	MDCT (8, 64, 128, 256 channel) with quad-phase CE, 5.0 mm thickness	Pathology	Yes	Size, WO, EC, LR
Ludwig DR et al (2019)	Retrospective	Cohort	Selective	America	131	65.2 (28–88)	131	HCC (27), NHM (104)	MDCT (64, 128 channel) with triple-phase CE, 2.0–3.0 mm thickness	Pathology	Yes	Size, APHE, WO, EC, LR
Sevim S et al (2019)	Retrospective	Cohort	Consecutive	Turkey	37	58.8 (NA)	37	NA	MDCT (16, 64 channel) with triple-phase CE, 2.0 mm thickness	Pathology	Yes	Size, APHE, WO, EC, LR
Zhang YD et al (2016)	Retrospective	Cohort	Consecutive	China	203	56.4 (31–83)	281	HCC (208), benign (73)	MDCT (16, 128 channel) with triple-phase CE, 1.2 mm thickness	Pathology or CCRS	Yes	Size, APHE, WO, EC

Articles are listed alphabetically according to the order of the names of the first authors

NA not available, *HCC* hepatocellular carcinoma, *NHM* non-HCC malignancy, *MDCT* multidetector CT, *CE* contrast enhancement, *CCRS* composite clinical reference standard, *APHE* arterial-phase hyperenhancement, *WO* nonperipheral washout, *EC* enhancing capsule, *LR* LI-RADS categorization

*Determined according to whether the readers were blinded to the final diagnoses of the analyzed lesions or not

Table 2 Imaging analysis methodologies of the included studies

Author (year of publication)	No. of readers	Years of experience	LI-RADS version	Presumed year of reading session	Size, ICC (95% CI)	APHE, κ (95% CI)	WO, κ (95% CI)	EC, κ (95% CI)	LR, κ (95% CI)
Alhasan A et al (2019)	2	R1: 23 years in liver imaging R2: 21 years in liver imaging	2017	2017	NA	0.90 (0.80–1.00)	0.93 (0.85–1.00)	0.66 (0.47–0.84)	0.90 (0.83–0.97)
Basha MAA et al (2018)	2	R1: 13 years in liver imaging R2: 10 years in liver imaging	2014	2017	NA	0.62 (0.37–0.87)	0.55 (0.28–0.82)	0.55 (0.18–0.92)	0.90 (0.79–1.00)
Cha DI et al (2017)	3	R1: 19 years in abdominal imaging R2: 8 years in abdominal imaging R3: 7 years in abdominal imaging	2014	2016	NA	0.62 (0.58–0.66)	0.84 (0.80–0.88)	0.80 (0.75–0.84)	NA
Chen N et al (2016)	2	Not reported	2014	2014	NA	NA	NA	NA	0.68 (0.59–0.77)
Chernyak V et al (2018)	2	R1: 10 years in abdominal imaging R2: 7 years in abdominal imaging	2014	2016	0.99 (0.99–1.00)	0.85 (0.65–1.00)	0.83 (0.65–1.00)	0.86 (0.72–1.00)	0.79 (0.68–0.88)
Ehman EC et al (2016)	2	All readers were board-certified abdominal radiologists	2014	2015	0.92 (0.88–0.95)	NA	NA	NA	NA
Fowler KJ et al (2018)	113	87%: specialist in liver imaging 12%: some expertise in liver imaging 1%: no expertise in liver imaging	2014	2015	NA	0.86 (0.82–0.91)	0.84 (0.80–0.90)	0.79 (0.72–0.85)	0.67 (0.61–0.71)
Joo I et al (2017)	2	R1: 10 years in abdominal imaging R2: 7 years in abdominal imaging	2014	2015	0.99 (0.98–0.99)	0.69 (0.54–0.84)	0.56 (0.38–0.75)	0.64 (0.54–0.75)	NA
Kim SS et al (2019)	2	R1: 23 years in abdominal imaging R2: 8 years in abdominal imaging	2017	2017	1.00 (1.00–1.00)	NA	0.45 (0.19–0.70)	0.48 (0.32–0.64)	0.66 (0.51–0.80)
Ludwig DR et al (2019)	2	R1: 7 years post-fellowship R2: 3 years post-fellowship	2018	2018	0.74 (0.66–0.81)	0.64 (0.47–0.80)	0.62 (0.44–0.80)	0.63 (0.44–0.82)	0.53 (0.29–0.77)
Sevim S et al (2019)	5	R1–R5: > 5 years in radiology experience	2014	2016	0.83 (0.74–0.91)	0.48 (0.38–0.58)	0.34 (0.23–0.44)	0.41 (0.31–0.52)	0.44 (0.39–0.49)
Zhang YD et al (2016)	2	R1: 6 years post-fellowship R2: 2 years post-fellowship	2014	2015	0.99 (0.99–0.99)	0.56 (0.47–0.65)	0.63 (0.49–0.77)	0.58 (0.53–0.63)	NA

LI-RADS Liver Imaging Reporting and Data System, *ICC* intraclass correlation coefficient, *κ* kappa value, *APHE* arterial-phase hyperenhancement, *WO* nonperipheral washout, *EC* enhancing capsule, *LR* LI-RADS categorization, *CI* confidence interval, *NA* not available

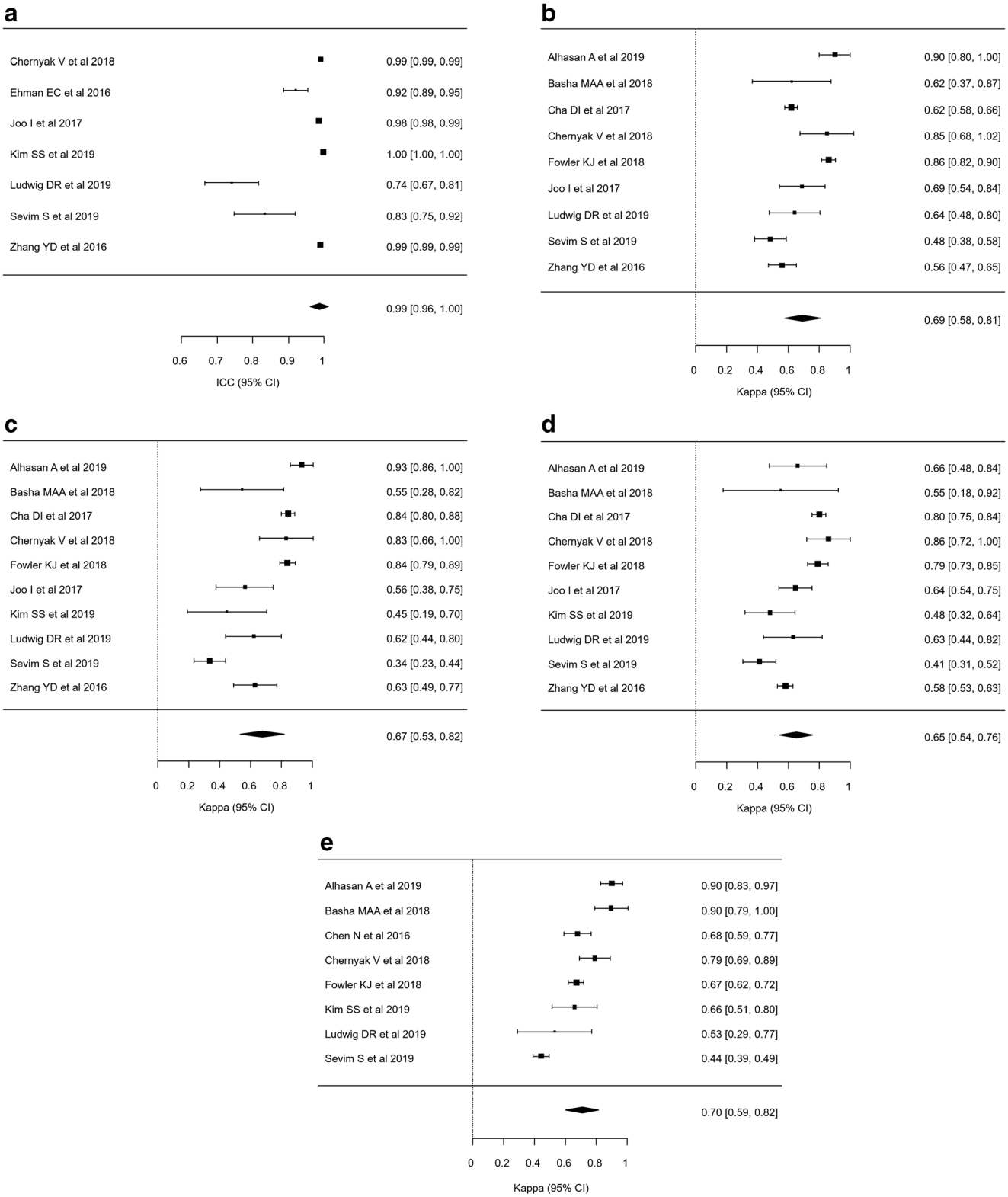

Fig. 2 Meta-analytic pooled inter-reader reliability for the CT Liver Imaging Reporting and Data System (LI-RADS). **a** Lesion size, **b** arterial-phase hyperenhancement, **c** nonperipheral washout, **d** enhancing capsule, and **e** LI-RADS categorization. ICC, intraclass correlation coefficient

size, APHE, WO, and EC, showed no significant differences in inter-reader reliability between studies with readers with ≥ 10 years of experience and those with < 10 years of experience ($p \geq 0.07$). In addition, the meta-analytic pooled κ for APHE and LI-RADS categorization in studies including inexperienced readers (< 5 years

Table 3 Subgroup analysis of inter-reader reliability in major features and LI-RADS categorizations according to the imaging analysis methodology

Covariates	Subgroup	Size, ICC		APHE, κ		WO, κ		EC, κ		LR, κ	
		Estimates (95% CI)	*p* value	Estimates (95% CI)	*p* value	Estimates (95% CI)	*p* value	Estimates (95% CI)	*p* value	Estimates (95% CI)	*p* value
No. of readers	Two readers	0.99 (0.96–1.00)	0.35	0.71 (0.57–0.86)	0.62	0.67 (0.51–0.83)	0.94	0.64 (0.53–0.75)	0.70	0.56 (0.33–0.78)	0.11
	≥ 3 readers	0.83 (0.74–0.91)		0.66 (0.44–0.87)		0.68 (0.35–1.00)		0.67 (0.42–0.92)		0.64 (0.43–0.85)	
Average reader experience*	≥ 10 years	0.99 (0.98–1.00)	0.56	0.74 (0.57–0.90)	0.07	0.74 (0.54–0.94)	0.14	0.69 (0.54–0.84)	0.13	0.82 (0.72–0.92)	0.01
	< 10 years	0.86 (0.71–1.00)		0.55 (0.47–0.62)		0.52 (0.32–0.72)		0.53 (0.41–0.66)		0.45 (0.39–0.50)	
Difference in reader experience†	All experienced readers	0.99 (0.98–1.00)	0.56	0.76 (0.63–0.90)	0.04	0.78 (0.63–0.93)	0.06	0.72 (0.60–0.84)	0.06	0.79 (0.68–0.89)	0.02
	Multiple readers with inexperienced reader	0.86 (0.71–1.00)		0.55 (0.47–0.62)		0.52 (0.32–0.72)		0.53 (0.41–0.66)		0.45 (0.39–0.50)	

The results were obtained using a random-effects model with or without Knapp and Hartung adjustment

LI-RADS Liver Imaging Reporting and Data System, *ICC* intraclass correlation coefficient, *APHE* arterial-phase hyperenhancement, *WO* nonperipheral washout, *EC* enhancing capsule, *LR* LI-RADS categorization, *CI* confidence interval

*The studies of Chen et al, Ehman et al, and Fowler et al were excluded as the average reader experience could not be estimated

† The studies of Chen et al and Ehman et al were excluded as they did not report the exact experience of each reader

of post-fellowship experience) were significantly lower than those in studies where all the readers were experienced (κ = 0.55 vs. 0.76, p = 0.04 and κ = 0.45 vs. 0.79, p = 0.02, respectively). Although studies including inexperienced readers showed lower meta-analytic pooled κ for WO (κ = 0.52 vs. 0.78) and EC (κ = 0.53 vs. 0.72) than studies with only experienced readers, the differences showed only borderline significance (p = 0.06 for both).

Subgroup analysis according to LI-RADS version

All LI-RADS versions including v2014, v2017, and v2018 showed substantial inter-reader reliability for the three major features (Table 4). For the LI-RADS categorization, LI-RADS v2018 had a pooled κ of 0.53, which was lower than that of v2014 (κ = 0.69) or v2017 (κ = 0.79). Regarding the presumed year of reading session, 88% (7/8) studies with LI-RADS v2014 were conducted after 2014, but the study with LI-RADS v2018 was conducted in 2018 (Table 2).

Meta-regression analysis

Substantial study heterogeneity was noted in all five variables of lesion size, APHE, WO, EC, and LI-RADS categorization ($I^2 \geq 90.0\%$ and $p < 0.001$). In the meta-regression analysis, clarity of blinding to the reference standard during review was significantly associated with study heterogeneity ($p \leq 0.04$; Supplementary Table 2). Studies with clear clarity of blinding showed substantial inter-reader reliability for APHE (κ = 0.64) and WO (κ = 0.62), but studies with unclear clarity of blinding showed almost perfect reliability (κ = 0.86 for APHE and 0.84 for WO). Regarding the CT technique, the number of detectors in MDCT showed a marginal significance with respect to lesion size (p = 0.05). However, other covariates including study design, study type, subject enrollment, multiphase CT, and slice thickness were not significant factors affecting study heterogeneity.

There was no significant publication bias with respect to lesion size, the three major features, or LI-RADS categorization ($p \geq 0.15$, Supplementary Figure 2).

Discussion

In this study, CT LI-RADS demonstrated substantial overall inter-reader reliability for major features and LI-RADS categorization, with meta-analytic pooled κ of 0.65–0.71. Substantial heterogeneity was noted in inter-reader reliability. The imaging analysis methodology varied across the studies, and the inter-reader reliability of CT LI-RADS differed significantly according to the average reader experience (p = 0.01) and the difference in reader experience (p = 0.02).

The meta-analytic κ for CT LI-RADS categorizations found in this study was similar to that in a previous study by Fowler et al (κ = 0.71 vs. 0.73) [9]. Considering the fact that Fowler et al performed a multi-center international study using a large number of readers and a mixture of all LI-RADS category assignments [9], their results would reflect those of clinical practice with minimal potential bias. However, the inter-reader reliabilities for major features found in the current study were lower than those in the previous study (κ = 0.65–0.69 vs. 0.84–0.88). This difference may have been due to differences in reader experience, i.e., the inclusion of inexperienced readers (< 5 years of post-fellowship experience) [12, 13, 31]. In the subgroup analysis of studies including all experienced readers, which excluded three studies that included inexperienced readers [12, 13, 31], the meta-analytic κ for major features was 0.72–0.78. Because the multi-center multi-reader study by Fowler et al was conducted in the year of 2014 and the three studies with inexperienced readers

Table 4 Subgroup analysis according to LI-RADS version

	Size, ICC (95% CI)	APHE, κ (95% CI)	WO, κ (95% CI)	EC, κ (95% CI)	LR, κ (95% CI)
LI-RADS v2014	0.98 (0.96–1.00)	0.67 (0.53–0.80)	0.66 (0.48–0.85)	0.67 (0.52–0.82)	0.69 (0.54–0.80)
I^2 statistics (%)	88.2	93.6	93.8	92.8	96.5
LI-RADS v2017	1.00 (1.00–1.00)	0.90 (0.80–1.00)	0.70 (0.23–1.00)	0.56 (0.39–0.74)	0.79 (0.55–1.00)
I^2 statistics *(%)*	0.0	0.0	92.1	50.7	88.5
LI-RADS v2018	0.74 (0.66–0.81)	0.64 (0.47–0.80)	0.62 (0.44–0.80)	0.63 (0.44–0.82)	0.53 (0.29–0.77)
I^2 statistics (%)	0.0	0.0	0.0	0.0	0.0

The results were obtained using a random-effects model with or without Knapp and Hartung adjustment

LI-RADS Liver Imaging Reporting and Data System, *ICC* intraclass correlation coefficient, *APHE* arterial-phase hyperenhancement, *κ* kappa value, *WO* nonperipheral washout, *EC* enhancing capsule, *LR* LI-RADS categorization, *CI* confidence interval

indicating the remained variability of CT LI-RADS were published after this multi-center multi-reader study, it may still be a problem requiring a solution. Therefore, to promote standardization and reproducibility of CT LI-RADS, a well-designed methodology for imaging analysis is necessary, and continuous education and updates for inexperienced readers are important.

In this meta-analysis, reader experience was one of the important factors affecting the inter-reader reliability of CT LI-RADS. The reported effects of reader experience on inter-reader reliability of LI-RADS are conflicting, with Davenport et al showing that experts had higher inter-reader reliability than novices [32], but Fowler et al finding that the inter-reader reliability of LI-RADS was not significantly associated with reader experience [9]. Considering these previous results and this meta-analysis together, reader experience may be one important factor associated with the inter-reader reliability of LI-RADS. In addition, our study showed that LI-RADS v2018 had relatively lower inter-reader reliability for LI-RADS categorization than LI-RADS v2014 or v2017. Because LI-RADS v2018 has recently been updated with a simplified definition of threshold growth and simplified criteria for LR-5 [2], these updates might not yet be familiar to radiologists. Although Fowler et al reported that LI-RADS inter-reader reliability was not significantly affected by LI-RADS familiarity, this previous study evaluated only LI-RADS v2014, and it could be difficult to evaluate the effect of LI-RADS version on inter-reader reliability. Considered together, these results indicate that the familiarity of radiologists with the latest version can also be an important factor influencing inter-reader reliability. To improve the relatively low inter-reader reliability of inexperienced readers and that associated with recently updated versions, education programs including training sets for young radiologists and regular feedback, and a comprehensive manual with both schematic figures and clinical examples to illustrate the features, would be helpful [33, 34].

Generally, as MRI provides multiparametric information from complex MRI sequences, MRI might be expected to have a lower inter-reader reliability for LI-RADS categorization than CT. However, the meta-analytic κ of CT LI-RADS categorizations in this study was very similar to that of MRI LI-RADS categorizations in a previous study (κ = 0.71 vs. 0.70) [14]. In addition, the inter-reader reliability of major features on CT was similar to that on MRI (APHE, κ = 0.69 on CT vs. 0.72 on MRI; WO, κ = 0.67 on CT vs. 0.69 on MRI; and EC, κ = 0.65 on CT vs. 0.66 on MRI) [14]. As LI-RADS uses the same definitions for major imaging features on CT and MRI [2], comparable inter-reader reliability might be expected between CT and MRI.

Our study showed that clarity of blinding to the reference standard during review was a significant factor associated with study heterogeneity, i.e., studies with unclear clarity of blinding to the reference standard showing higher inter-reader reliability for APHE (0.86 vs. 0.64, $p = 0.03$) and WO (0.84 vs. 0.62, $p = 0.04$) than those with clear clarity of blinding. Although inter-reader reliability can be evaluated according to the agreement between readers without considering the reference standard, our result suggests that knowledge of the final diagnosis might affect inter-reader reliability. Considering the fact that a recent study of MRI LI-RADS inter-reader reliability reported a similar result [14], further study is needed to evaluate why the clarity of blinding to the reference standard is associated with inter-reader reliability.

This study has some limitations. First, study heterogeneity and variations in imaging analysis methodologies were noted. To explore the causes of the variations, we robustly performed subgroup analyses and meta-regression analyses. Second, we could not include 10 articles because of insufficient data for determining inter-reader reliability. As the meta-analysis of inter-reader reliability required the standard variance as well as the ICC or κ from each study, studies that provided only the ICC or κ without standard variance were excluded. Third, although a recent meta-analysis reported that the inter-reader reliability of MRI LI-RADS, i.e., the pooled kappa value of LI-RADS categorization was 0.70 [14], our study provides additional information in that it reports the inter-reader reliability of CT LI-RADS and differences in it according to imaging analysis methodology, which were not covered in the previous study.

In conclusion, CT LI-RADS demonstrated substantial inter-reader reliability for major features and LI-RADS categorizations. The imaging analysis methodology varied across studies, and the inter-reader reliability of CT LI-RADS differed significantly according to the average reader experience and the difference in reader experience. Therefore, the reported results for inter-reader reliability of CT LI-RADS in the literature should be understood with consideration of the imaging analysis methodology.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00330-021-07815-y.

Funding This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (grant number: NRF-2019R1G1A1099743).

Compliance with ethical standards

Guarantor The scientific guarantor of this publication is Sang Hyun Choi.

Conflict of interest The authors of this manuscript declare no relationships with any companies whose products or services may be related to the subject matter of the article.

Statistics and biometry One of the authors has significant statistical expertise (Ji Sung Lee).

Informed consent Written informed consent was not required for this study because this is a meta-analysis.

Ethical approval Institutional Review Board approval was not required because this is a meta-analysis.

Methodology

- Systematic Review and Meta-analysis

References

1. Mitchell DG, Bruix J, Sherman M, Sirlin CB (2015) LI-RADS (Liver Imaging Reporting and Data System): summary, discussion, and consensus of the LI-RADS Management Working Group and future directions. Hepatology 61:1056–1065
2. American College of Radiology CT/MRI LI-RADS v2018. Available via https://www.acr.org/Clinical-Resources/Reporting-and-Data-Systems/LI-RADS/CT-MRI-LI-RADS-v2018. Accessed May 27, 2019.
3. Chernyak V, Fowler KJ, Kamaya A et al (2018) Liver Imaging Reporting and Data System (LI-RADS) Version 2018: imaging of hepatocellular carcinoma in at-risk patients. Radiology 289:816–830
4. Marrero JA, Kulik LM, Sirlin CB et al (2018) Diagnosis, Staging, and Management of Hepatocellular Carcinoma: 2018 Practice Guidance by the American Association for the Study of Liver Diseases. Hepatology 68:723–750
5. Clavien PA, Lesurtel M, Bossuyt PM et al (2012) Recommendations for liver transplantation for hepatocellular carcinoma: an international consensus conference report. Lancet Oncol 13:e11–e22
6. Kim DH, Choi SH, Park SH et al (2019) Meta-analysis of the accuracy of Liver Imaging Reporting and Data System category 4 or 5 for diagnosing hepatocellular carcinoma. Gut 68:1719–1721
7. Lee SM, Lee JM, Ahn SJ, Kang HJ, Yang HK, Yoon JH (2019) LI-RADS Version 2017 versus Version 2018: diagnosis of hepatocellular carcinoma on gadoxetate disodium-enhanced MRI. Radiology 292:655–663
8. van der Pol CB, Lim CS, Sirlin CB et al (2019) 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 156:976–986
9. Fowler KJ, Tang A, Santillan C et al (2018) Interreader reliability of LI-RADS Version 2014 algorithm and imaging features for diagnosis of hepatocellular carcinoma: a large international multireader study. Radiology 286:173–185
10. Alhasan A, Cerny M, Olivie D et al (2019) LI-RADS for CT diagnosis of hepatocellular carcinoma: performance of major and ancillary features. Abdom Radiol (NY) 44:517–528
11. Kim SS, Hwang JA, Shin HC et al (2019) LI-RADS v2017 categorisation of HCC using CT: Does moderate to severe fatty liver affect accuracy? Eur Radiol 29:186–194
12. Ludwig DR, Fraum TJ, Cannella R et al (2019) Expanding the Liver Imaging Reporting and Data System (LI-RADS) v2018 diagnostic population: performance and reliability of LI-RADS for distinguishing hepatocellular carcinoma (HCC) from non-HCC primary liver carcinoma in patients who do not meet strict LI-RADS high-risk criteria. HPB (Oxford) 21:1697–1706
13. Sevim S, Dicle O, Gezer NS, Baris MM, Altay C, Akin IB (2019) How high is the inter-observer reproducibility in the LIRADS reporting system? Pol J Radiol 84:e464–e469
14. Kang JH, Choi SH, Lee JS et al (2020) Interreader agreement of Liver Imaging Reporting and Data System on MRI: a systematic review and meta-analysis. J Magn Reson Imaging 52:795–804
15. Brady A, Laoide RO, McCarthy P, McDermott R (2012) Discrepancy and error in radiology: concepts, causes and consequences. Ulster Med J 81:3–9
16. Brady AP (2017) Error and discrepancy in radiology: inevitable or avoidable? Insights Imaging 8:171–182
17. Stroup DF, Berlin JA, Morton SC et al (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis Of Observational Studies in Epidemiology (MOOSE) group. JAMA 283:2008–2012
18. Liberati A, Altman DG, Tetzlaff J et al (2009) The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate healthcare interventions: explanation and elaboration. BMJ 339: b2700
19. Moher D, Liberati A, Tetzlaff J, Altman DG, Group P (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. BMJ 339:b2535
20. Tang A, Hallouch O, Chernyak V, Kamaya A, Sirlin CB (2018) Epidemiology of hepatocellular carcinoma: target population for surveillance and diagnosis. Abdom Radiol (NY) 43:13–25
21. Kottner J, Audige L, Brorson S et al (2011) Guidelines for Reporting Reliability and Agreement Studies (GRRAS) were proposed. J Clin Epidemiol 64:96–106
22. IntHout J, Ioannidis JP, Borm GF (2014) The Hartung-Knapp-Sidik-Jonkman method for random effects meta-analysis is straightforward and considerably outperforms the standard DerSimonian-Laird method. BMC Med Res Methodol 14:25
23. Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. Biometrics 33:159–174
24. Higgins JP, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327:557–560
25. Basha MAA, AlAzzazy MZ, Ahmed AF et al (2018) Does a combined CT and MRI protocol enhance the diagnostic efficacy of LI-RADS in the categorization of hepatic observations? A prospective comparative study. Eur Radiol 28:2592–2603
26. Cha DI, Jang KM, Kim SH, Kang TW, Song KD (2017) Liver Imaging Reporting and Data System on CT and gadoxetic acid-enhanced MRI with diffusion-weighted imaging. Eur Radiol 27: 4394–4405
27. Chen N, Motosugi U, Morisaka H et al (2016) Added value of a gadoxetic acid-enhanced hepatocyte-phase image to the LI-RADS system for diagnosing hepatocellular carcinoma. Magn Reson Med Sci 15:49–59
28. Chernyak V, Flusberg M, Law A, Kobi M, Paroder V, Rozenblit AM (2018) Liver Imaging Reporting and Data System: discordance between computed tomography and gadoxetate-enhanced magnetic resonance imaging for detection of hepatocellular carcinoma major features. J Comput Assist Tomogr 42:155–161
29. Ehman EC, Behr SC, Umetsu SE et al (2016) Rate of observation and inter-observer agreement for LI-RADS major features at CT and MRI in 184 pathology proven hepatocellular carcinomas. Abdom Radiol (NY) 41:963–969
30. Joo I, Lee JM, Lee DH, Ahn SJ, Lee ES, Han JK (2017) Liver imaging reporting and data system v2014 categorization of hepatocellular carcinoma on gadoxetic acid-enhanced MRI: comparison with multiphasic multidetector computed tomography. J Magn Reson Imaging 45:731–740
31. Zhang YD, Zhu FP, Xu X et al (2016) Classifying CT/MR findings in patients with suspicion of hepatocellular carcinoma: comparison of liver imaging reporting and data system and criteria-free Likert scale reporting models. J Magn Reson Imaging 43:373–383

32. Davenport MS, Khalatbari S, Liu PS et al (2014) Repeatability of diagnostic features and scoring systems for hepatocellular carcinoma by using MR imaging. Radiology 272:132–142
33. Chernyak V, Sirlin CB (2020) Editorial for "Interreader Agreement of Liver Imaging Reporting and Data System on MRI: A Systematic Review and Meta Analysis". J Magn Reson Imaging 52:805–806
34. Curci NE, Gartland P, Shankar PR et al (2018) Long-distance longitudinal prostate MRI quality assurance: from startup to 12 months. Abdom Radiol (NY) 43:2505–2512