Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Interobserver agreement of LR-TRA 2017 and 2024 and mRECIST criteria on MRI after locoregional therapy: Impact of ancillary features and reader experience

Alba Salgado-Parente

albasalgado9@gmail.com

Sunnybrook Health Science Centre

Ana Villanueva Campos

Hospital Universitario San Rafael

Elena Canales Lachén

Hospital Universitario Ramón y Cajal

Luis González Campo

Hospital San Pedro

Abel González Huete

Hospital Universitario Ramón y Cajal

Diego A García Paredes

Hospital Universitario Ramón y Cajal

Roberto Cannella

University of Palermo

Research Article

Keywords: Liver Neoplasms, Carcinoma, Hepatocellular, Observer Variation, Catheter Ablation, Chemoembolization, Therapeutic, Treatment Outcome

Posted Date: November 19th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11614816/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Interobserver agreement of LR-TRA 2017 and 2024 and mRECIST criteria on MRI after locoregional therapy: Impact of ancillary features and reader experience

Abstract

Purpose: To compare interobserver agreement of the Liver Imaging Reporting and Data System Treatment Response Algorithm (LR-TRA) v2017 and v2024, with and without ancillary features, and of modified Response Evaluation Criteria in Solid Tumors (mRECIST) for hepatocellular carcinoma (HCC) after nonradiation locoregional therapy, and to assess the effects of reader experience, ancillary features, and treatment type.

Methods: In this retrospective, single-center study, six radiologists, four of them subspecialized in abdominal imaging, independently classified 100 treated HCC lesions on MRI using LR-TRA v2017, v2024 without and with ancillary features, and mRECIST. Agreement was assessed with Fleiss' kappa globally and by category, reader group, and treatment type.

Results: Global agreement was moderate for LR-TRA v2017 ($\kappa = 0.508$), v2024 without ($\kappa = 0.538$) and with ancillary features ($\kappa = 0.554$), and substantial for mRECIST ($\kappa = 0.727$). Among subspecialized readers, agreement increased for all LR-TRA variants ($\kappa = 0.639–0.646$; $p \leq 0.016$) but not for mRECIST ($p = 0.966$). The equivocal category was the least reproducible and the most dependent on reader experience (all $p \leq 0.025$). Ancillary features reduced equivocal classifications by 44.4%, with restricted diffusion present in 89.3% of reclassified observations, although agreement on reclassification was low ($\kappa = 0.281$). A treatment-related difference in agreement observed with v2017 was absent with v2024.

Conclusion: LR-TRA v2024 showed numerically higher interobserver agreement than v2017, but reproducibility remained moderate. Reader experience had a greater influence on agreement than the LR-TRA version, particularly for the equivocal category. Ancillary features reduced equivocal classifications without improving agreement within this category.

Key Words: Liver Neoplasms; Carcinoma, Hepatocellular; Observer Variation; Catheter Ablation; Chemoembolization, Therapeutic; Treatment Outcome

Introduction

Hepatocellular carcinoma (HCC) is the most common primary liver cancer and remains a leading cause of cancer-related death worldwide (1). Locoregional therapy (LRT), including radiofrequency ablation and transarterial chemoembolization, is widely used with definitive intent, as a bridge or downstaging strategy before liver transplantation, and for palliation. Because HCC is frequently multifocal and may develop metachronously, treatment response is commonly assessed on a per-lesion basis. A reproducible and reliable response assessment system is therefore essential to guide subsequent management.

The CT/MRI Liver Imaging Reporting and Data System (LI-RADS) Treatment Response Algorithm (LR-TRA) was introduced in 2017 to provide a standardized framework for assessing the viability of treated HCC, classifying lesions as nonviable, viable, equivocal, or nonevaluable (2). Overall interobserver agreement with LR-TRA v2017 has generally been reported as moderate to substantial (2–4), but this headline figure obscures meaningful differences between categories: agreement for the equivocal category in particular has been consistently lower, ranging from only fair to substantial across studies (4–7). This variability is clinically relevant: a substantial proportion of lesions categorized as equivocal after nonradiation LRT demonstrate incomplete necrosis at pathologic examination (5,8), with direct implications for decisions about retreatment and surveillance.

The 2024 LR-TRA update introduced several important modifications aimed at improving the assessment of treated HCC. For the first time, separate algorithms were established for nonradiation- and radiation-based LRT, reflecting the different post-treatment response patterns and imaging appearances associated with these treatment modalities. For nonradiation-treated HCC, the revised algorithm simplified the definition of viable tumor and introduced the optional use of two ancillary MRI features (restricted diffusion and mild-to-moderate T2 hyperintensity) to allow lesions initially considered equivocal to be upgraded to viable (Figure 1). Improving reader confidence and concordance was an explicit goal of the 2024 revision (2,9).

This updated classification has already been shown to improve diagnostic performance, with higher sensitivity and accuracy when ancillary features are incorporated (9,10). However, interobserver agreement with v2024 has so far been reported only in studies with two or three readers (9,11), and it remains unclear whether the diagnostic gain is accompanied by improved reproducibility, particularly for the equivocal category, and whether it is consistent across readers with different levels of experience, a factor that has been shown to meaningfully influence interobserver agreement in the closely analogous Prostate Imaging Reporting and Data System (PI-RADS) (12,13).

The purpose of this study was therefore to compare the interobserver agreement of LR-TRA v2017, v2024 with and without ancillary features, and modified Response Evaluation Criteria in Solid Tumors (mRECIST) for assessing HCC response after nonradiation locoregional therapy,

and to determine how that agreement is affected by the addition of ancillary features, by reader experience, and by treatment type, with particular attention to the equivocal category.

Materials and Methods

1. Study design and patient population

This retrospective, single-center study was approved by the institutional review board, which waived the requirement for informed consent given its retrospective design. Consecutive patients treated for HCC with LRT between July 2019 and January 2025 were screened for inclusion.

Lesions were included if they met all of the following criteria: (a) observation categorized as LR-4 or LR-5 on pretreatment CT or MRI and treated as HCC; (b) treatment with radiofrequency ablation (RFA), transarterial chemoembolization (TACE), or combined therapy (sequential RFA and TACE for the same lesion within a single treatment strategy); and (c) follow-up MRI performed within 1 year after treatment. When multiple follow-up MRIs were available within this window, the examination closest to treatment was analyzed.

Lesions were excluded if any of the following applied: (a) prior or concurrent radiotherapy or radioembolization, (b) prior or concurrent systemic therapy, (c) follow-up MRI performed without intravenous contrast material, or (d) follow-up MRI images unavailable for review for technical reasons.

In patients with multiple concurrently treated lesions, the largest lesion was designated as the target lesion. Metachronous lesions (new lesions treated in a subsequent year) were included as additional independent cases. The lesion selection process is summarized in Figure 2.

This yielded a final cohort of 97 patients (83 men, 14 women; mean age 64.3 ± 11.1 years, median 64, range 20–88), contributing 100 treated lesions (three patients had two treated lesions each). Lesions were treated with RFA (n = 74), TACE (n = 16), or combined therapy (n = 10). Twelve lesions (12%) had received prior locoregional treatment before the session evaluated in this study; the remaining 88 (88%) were treated for the first time. No formal sample size calculation was performed; the final cohort comprised all consecutive patients meeting the eligibility criteria during the study period.

2. Imaging protocol

All patients underwent contrast-enhanced MRI on either a 1.5-T scanner (Ingenia, Achieva, or Achieva dStream; Philips Healthcare) or a 3.0-T scanner (Ingenia Elition X or MR 7700, Philips Healthcare; SIGNA Premier, GE HealthCare) after intravenous administration of gadobutrol (Gadovist; Bayer HealthCare, Berlin, Germany). The median interval between treatment and MRI was 98 days (interquartile range, 42–170). The imaging protocol included fat-suppressed T1-weighted sequences acquired before and dynamically after intravenous contrast administration, including the late arterial phase (approximately 35–40 seconds), portal venous phase

(approximately 60–70 seconds), and delayed phase (approximately 3–5 minutes) after contrast injection. Additional sequences included T2-weighted turbo spin-echo sequences and diffusion-weighted imaging (high b-value, 800 s/mm²) with corresponding apparent diffusion coefficient (ADC) maps. Subtraction images were not routinely generated at the study institution and were therefore not evaluated.

3. Image analysis

Six radiologists with varying levels of clinical experience in liver imaging independently and blindly reviewed all MRI examinations on a Synapse picture archiving and communication system (PACS) workstation (version 7.4.100; Fujifilm), without access to clinical history, treatment details, or other readers' assessments. Readers were stratified into two groups: four subspecialized abdominal radiologists with dedicated expertise in abdominopelvic imaging, and two non-specialized readers (one general radiologist and one resident). The four subspecialized readers had 10, 8, 7, and 6 years of experience; the general radiologist and resident had 6 and 3 years, respectively. A formal calibration session was performed prior to the reading sessions.

All six readers independently classified each lesion according to four systems, yielding a total of 2,400 readings (6 readers × 100 lesions × 4 systems). To minimize recall bias, each system was evaluated in a separate reading session, with an interval of at least 2 months between sessions, and readers had no access to their classifications from previous sessions. Pretreatment imaging was available for comparison. The four classification systems were evaluated in the following order, and the order of cases within each session was fixed and identical for all readers:

- LR-TRA v2017;
- LR-TRA v2024, without ancillary features;
- LR-TRA v2024, with ancillary features; readers additionally recorded the presence or absence of the two ancillary features (T2 hyperintensity and restricted diffusion on DWI) for all lesions;
- mRECIST: although mRECIST was originally developed for patient-level response assessment, it was applied here at the lesion level, consistent with the lesion-level design of the other three systems.

For each LR-TRA system, lesions were categorized as LR-TR nonviable, equivocal, viable, or nonevaluable. For mRECIST, lesions were categorized according to standard mRECIST response categories: complete response, partial response, stable disease, or progressive disease.

Readings were recorded prospectively in REDCap (Research Electronic Data Capture) (14), a web-based software platform in which each patient was identified by a pseudonymized code, in compliance with the General Data Protection Regulation (GDPR).

4. Statistical analysis

Interobserver agreement was assessed using Fleiss' kappa (κ) for multi-rater agreement, calculated at two levels:

- Global agreement, for each of the four classification systems (LR-TRA v2017, v2024 without ancillary features, v2024 with ancillary features, and mRECIST)
- Category-specific agreement, calculated separately for each category within each system (nonviable, equivocal, viable, and nonevaluable for LR-TRA; complete response, partial response, stable disease, and progressive disease for mRECIST)

Kappa values were interpreted according to Landis and Koch criteria (15): ≤0.00 poor, 0.01–0.20 slight, 0.21–0.40 fair, 0.41–0.60 moderate, 0.61–0.80 substantial, and 0.81–1.00 almost perfect agreement. Ninety-five percent confidence intervals (CIs) for kappa were calculated using the large-sample variance formula of Fleiss, Levin, and Paik (16).

The influence of reader experience was examined by repeating the analysis in the subgroup of the four subspecialized abdominal radiologists and compared with the full six-reader panel. Because this subgroup was nested within the full panel, both global and category-specific comparisons used a paired bootstrap (5,000 resamples, resampling patients so that lesions from the same patient remained together) rather than an independent-samples test.

To assess the impact of treatment type, the analysis was repeated separately in lesions treated with RFA (n = 74) and in lesions treated with TACE or combined therapy (n = 26). These two groups were pooled because of their small individual sizes and because both share embolization-related posttreatment imaging features, such as heterogeneous necrosis and altered arterial enhancement, which differ from the more uniform appearance of the ablation zone after RFA. Because these were independent, non-overlapping groups, differences in κ were tested with a Z-statistic (SE = $\sqrt{[2(1-\kappa)^2 / (n(k-1))]}$), with $p < 0.05$ considered significant. No correction for multiple comparisons was applied, and category-specific and subgroup analyses were considered exploratory.

The effect of ancillary features was evaluated by examining category redistribution between readings without and with ancillary features. Interobserver agreement on whether a lesion changed category (reclassified vs. retained) was calculated using Fleiss' κ. Among lesions reclassified from equivocal to viable, the presence of each ancillary feature (T2 hyperintensity and restricted diffusion on DWI) was tabulated.

Statistical analyses were performed with Python version 3.12 (Python Software Foundation, Wilmington, DE, USA) using the statsmodels, scikit-learn, and SciPy libraries.

Given that some published studies report weighted rather than unweighted κ for these ordinal categories (7,11,17), global agreement for each LR-TRA system was additionally estimated as a sensitivity analysis using linear-weighted and quadratic-weighted Cohen's κ, averaged across all pairwise reader comparisons, for both the full six-reader panel and the subgroup of four

subspecialized readers. Lesions rated as nonevaluable by either reader in each pair were excluded from that pairwise comparison, because the nonevaluable category does not fit an ordinal viability scale.

Results

Interobserver agreement across classification systems

Global interobserver agreement among the six readers was moderate for all three LR-TRA classifications with $\kappa = 0.508$ (95% CI: 0.399–0.617) for LR-TRA v2017, $\kappa = 0.538$ (95% CI: 0.407–0.668) for LR-TRA v2024 without ancillary features, and $\kappa = 0.554$ (95% CI: 0.424–0.683) for LR-TRA v2024 with ancillary features. Differences between LR-TRA versions were small ($\Delta\kappa \leq 0.046$). Agreement was higher for mRECIST ($\kappa = 0.727$, 95% CI: 0.553–0.900), which fell in the substantial range.

Category-specific agreement followed a consistent pattern across the three LR-TRA systems: the nonviable ($\kappa = 0.57$–0.58) and viable ($\kappa = 0.60$–0.66) categories showed the highest agreement, whereas the equivocal ($\kappa = 0.30$–0.38) and nonevaluable ($\kappa = 0.29$–0.35) categories showed the lowest (Table 1, Figure 3). Agreement for the nonevaluable category should be interpreted with caution given its small size (n = 9). For mRECIST, agreement was substantial for complete response ($\kappa = 0.75$), stable disease ($\kappa = 0.76$), and partial response ($\kappa = 0.68$), whereas κ could not be estimated for progressive disease given its very low frequency (n = 1).

Impact of reader experience

Agreement was consistently higher among the subspecialized subgroup for all three LR-TRA classifications (Table 2, Figure 4). For LR-TRA v2017, agreement improved from moderate ($\kappa = 0.508$) to substantial ($\kappa = 0.639$; $p < 0.001$). A similar pattern was observed for LR-TRA v2024 without ancillary features ($\kappa = 0.538$ vs 0.643; $p = 0.004$) and with ancillary features ($\kappa = 0.554$ vs 0.646; $p = 0.016$). In contrast, reader experience had no measurable effect on mRECIST agreement ($\kappa = 0.727$ vs 0.726; $p = 0.966$).

The effect of reader experience was most pronounced in the LR-TR equivocal category. The improvement in agreement for this category was statistically significant across all three LR-TRA systems (all $p \leq 0.025$), with kappa increasing from 0.379 to 0.515 for LR-TRA v2017, from 0.372 to 0.529 for v2024 without ancillary features, and from 0.302 to 0.467 for v2024 with ancillary features among subspecialized readers (Figure 5). Improvement was also significant for the nonviable category in all three LR-TRA assessments, and for the viable category only with v2017. The nonevaluable category was not considered further because of its very small number of cases (n = 9).

Sensitivity analysis: weighted kappa

Weighted kappa estimates were higher than the unweighted Fleiss' kappa used as the primary metric across both reader groups, because weighting credits near-miss disagreements (Table S1).

For the full six-reader cohort, quadratic-weighted kappa ranged from 0.670 to 0.702 across the three LR-TRA assessments, placing agreement in the substantial range, compared with moderate agreement under the primary unweighted metric. For the four-reader subspecialized subgroup, quadratic-weighted kappa ranged from 0.737 to 0.808, reaching the almost-perfect range for LR-TRA v2017.

Impact of treatment type

Global interobserver agreement varied by treatment type for LR-TRA v2017, but not for LR-TRA v2024 or mRECIST (Table 3, Figure 6). For LR-TRA v2017, agreement was lower in lesions treated with RFA ($\kappa = 0.432$) than in lesions treated with TACE or combined therapy ($\kappa = 0.647$) ($p = 0.006$). This difference was no longer present for LR-TRA v2024, where agreement was similar between RFA and TACE/combined therapy. No significant difference by treatment type was observed for mRECIST.

Impact of ancillary features in LR-TRA v2024

Among the 63 reader-case observations initially classified as LR-TR equivocal without ancillary features, 28 (44.4%) were reclassified as viable after incorporating ancillary features (Table 4), while the remaining 35 (55.6%) remained equivocal.

Interobserver agreement on whether a given reader changed their classification (versus retained it) after applying ancillary features was low (Fleiss' $\kappa = 0.281$). Consistent with this, category-specific agreement for the equivocal category decreased slightly after incorporating ancillary features ($\kappa = 0.372$ without ancillary features vs. $\kappa = 0.302$ with ancillary features).

Restricted diffusion was present in most reclassified observations (89.3%), whereas T2 hyperintensity was present in fewer than half (46.4%) and rarely occurred in isolation (Table 4, Panel B).

Discussion

Our study showed numerically slightly higher interobserver agreement with LR-TRA v2024 than with v2017, both without ancillary features ($\kappa = 0.538$ vs. 0.508) and with ancillary features ($\kappa = 0.554$ vs. 0.508), although agreement for all three LR-TRA assessments remained lower than with mRECIST ($\kappa = 0.727$). Previously reported LR-TRA reproducibility, derived almost exclusively from studies applying v2017, is highly heterogeneous, with a pooled meta-analytic estimate of $\kappa = 0.70$ (95% CI, 0.58–0.82) across CT and MRI, and individual two-reader studies ranging from moderate ($\kappa = 0.58$) to almost perfect ($\kappa = 0.94$) agreement (Table S2) (18). Two previous studies have reported agreement for v2024 with small expert panels. Zhou et al. found comparable weighted agreement for v2017 ($\kappa = 0.73$) and v2024 without and with ancillary features ($\kappa = 0.72$ and 0.75) between two readers, with lower agreement for mRECIST ($\kappa = 0.67$) (11), and Shin et al., using hepatobiliary agent-enhanced MRI and three fellowship-trained abdominal radiologists, reported moderate Fleiss' κ for v2017 and v2024 (0.52 and 0.54), with higher agreement at CT

(0.65 for both versions) (9). Our unweighted estimates were close to those of Shin et al. but below the weighted values reported in most earlier studies. Part of this gap reflects the agreement metric: expressed as quadratic-weighted κ, as in several previous studies, agreement in our full panel (κ = 0.670–0.702) was close to the pooled estimate and to the values reported by Zhou et al. The remaining difference may partly reflect the composition of our six-reader panel (four subspecialists, one general radiologist, and one resident, with 3 to 10 years of experience). Because this panel more closely resembles real-world clinical practice than the small, calibrated expert panels used in previous studies, our results may provide a more representative estimate of LR-TRA reproducibility in routine clinical settings (6,7,11).

Supporting this interpretation, agreement increased significantly when the analysis was restricted to our four subspecialized readers (κ = 0.64–0.65; $p \leq 0.016$), consistent with the experience gradient described by Wang et al. (4). In this subgroup, agreement was nearly identical across the three LR-TRA assessments, indicating that reader experience had a greater influence on reproducibility than algorithm version. Because years of practice overlapped between groups, this effect is better attributed to subspecialization than to years of experience alone. Consistent with previous reports, the equivocal category was the least reproducible component of LR-TRA in our series across both v2017 and v2024 (4,6). The effect of reader experience was most evident in this category, where agreement increased from fair to moderate across all three LR-TRA assessments, suggesting that subspecialist expertise may be particularly helpful in these borderline, image-dependent cases.

In our cohort, incorporation of ancillary features reduced the proportion of equivocal reader-case observations by 44.4%, consistent with Zhou et al. (11) and Shin et al. (9), supporting the goal of LR-TRA v2024 of reducing diagnostic uncertainty. Nevertheless, agreement on the reclassification process remained low (κ = 0.281), likely indicating that ancillary features reduce diagnostic uncertainty numerically without fully resolving the underlying subjectivity. Shin et al. similarly attributed the lower agreement at hepatobiliary agent-enhanced MRI than at CT to the subjectivity and discretionary use of ancillary features (9). Restricted diffusion was by far the ancillary feature most frequently used to reclassify equivocal observations as viable (89.3% of reclassified cases), substantially more often than T2 hyperintensity, consistent with prior evidence showing greater specificity of DWI than T2 signal for viable tumor (90.0% vs. 56.8%) (19). This suggests that diffusion-based criteria may offer more consistent grounds for reclassification than the more subjective assessment of T2 signal.

Treatment type also influenced reproducibility. Agreement for LR-TRA v2017 was significantly lower in RFA than in TACE/combined-treated lesions (κ = 0.432 vs. 0.647, $p = 0.006$), a finding opposite to that reported in previous studies, in which ablation-treated lesions have generally demonstrated higher agreement or diagnostic performance than arterially treated lesions (8,20). This difference was no longer apparent with either v2024 variant ($p = 0.843$ and 0.981). These findings suggest that the effect of treatment type on LR-TRA reproducibility may be attenuated by

the v2024 framework, although the underlying reasons for the discrepancy with prior reports remain uncertain.

These findings have three practical implications. First, the choice between v2017 and v2024 had less influence on reproducibility than reader composition. Second, equivocal observations, the least reproducible category and the one most dependent on reader expertise, may benefit from subspecialist review when findings remain indeterminate. Third, although ancillary features reduced equivocal classifications, agreement on reclassification remained low, so their application would benefit from further standardization.

These findings should be interpreted in light of several study limitations. First, it was retrospective and single-center, and no formal sample size calculation was performed, which may limit generalizability and the precision of subgroup estimates. Second, treatment-type subgroups were unbalanced; however, patients were included consecutively, so this distribution reflects the real-world use of these therapies at our institution rather than a selection artifact. Third, the interval between treatment and follow-up MRI varied widely, which could have introduced heterogeneity in the imaging appearance of treated lesions across the cohort. Fourth, cross-study comparisons of kappa, should be interpreted cautiously, as kappa is sensitive to the prevalence of viable versus nonviable lesions and to reader-panel composition, both of which differed between cohorts.

Conclusion

In this multireader study, LR-TRA v2024 showed numerically slightly higher interobserver agreement than v2017 and reduced the frequency of equivocal classifications with the incorporation of ancillary features, although overall reproducibility remained moderate. Reader experience had a greater influence on reproducibility than the specific LR-TRA version, particularly for the equivocal category. These findings offer qualified support for the incremental value of the 2024 framework while highlighting the important role of subspecialist expertise in indeterminate post-treatment cases.

TABLES

Panel A. LR-TRA CRITERIA

Category	LR-TRA v2017	LR-TRA v2024, without ancillary features	LR-TRA v2024, with ancillary features
Nonviable	0.567 (0.468–0.666)	0.583 (0.461–0.705)	0.583 (0.461–0.705)
Equivocal	0.379 (0.193–0.566)	0.372 (0.160–0.584)	0.302 (−0.004–0.608)
Viable	0.603 (0.395–0.811)	0.663 (0.436–0.889)	0.652 (0.480–0.824)
Nonevaluable	0.290 (−0.154–0.734)[a]	0.346 (−0.295–0.986)[a]	0.346 (−0.295–0.986)[a]
Global	0.508 (0.399–0.617)	0.538 (0.407–0.668)	0.554 (0.424–0.683)

Panel B. mRECIST

mRECIST category	κ (95% CI)
Complete response	0.750 (0.582–0.917)
Stable disease	0.764 (0.463–1.000)[a]
Partial response	0.679 (0.435–0.922)
Progressive disease	not estimable (n = 1)
Global	0.727 (0.553–0.900)

Table 1. Fleiss' κ (95% CI) by category, for each LR-TRA version and for mRECIST. Agreement was interpreted according to conventional Landis–Koch criteria: poor/slight (κ < 0.21), fair (κ = 0.21–0.40), moderate (κ = 0.41–0.60), substantial (κ = 0.61–0.80), and almost perfect (κ = 0.81–1.00). [a] Wide confidence interval reflecting the small number of cases in this category.

Classification system	κ (6 readers), 95% CI	κ (4 subspecialized), 95% CI	Difference	95% CI of difference	p
LR-TRA v2017	0.508 (0.399–0.617)	0.639 (0.499–0.778)	+0.130	[+0.065, +0.202]	<0.001
LR-TRA v2024, without ancillary features	0.538 (0.407–0.668)	0.643 (0.493–0.793)	+0.105	[+0.031, +0.181]	0.004
LR-TRA v2024, with ancillary features	0.554 (0.424–0.683)	0.646 (0.496–0.795)	+0.092	[+0.017, +0.168]	0.016
mRECIST	0.727 (0.553–0.900)	0.726 (0.518–0.935)	0.000	[−0.066, +0.071]	0.966

Table 2. Global Fleiss' κ by classification system, comparing all readers with the subspecialized-reader subgroup. Agreement was interpreted according to conventional Landis–Koch criteria: poor/slight (κ < 0.21), fair (κ = 0.21–0.40), moderate (κ = 0.41–0.60), substantial (κ = 0.61–0.80), and almost perfect (κ = 0.81–1.00).

Classification system	κ, RFA (n = 74), 95% CI	κ, TACE/combined (n = 26), 95% CI	Z	p-value
LR-TRA v2017	0.432 (0.288–0.575)	0.647 (0.495–0.798)	2.75	0.006
LR-TRA v2024, without ancillary features	0.526 (0.353–0.699)	0.543 (0.361–0.725)	0.20	0.843
LR-TRA v2024, with ancillary features	0.548 (0.376–0.720)	0.546 (0.366–0.727)	−0.02	0.981
mRECIST	0.700 (0.453–0.948)	0.739 (0.526–0.952)	0.77	0.441

Table 3. Global Fleiss' κ by classification system, comparing RFA with TACE/combined therapy. Agreement was interpreted according to conventional Landis–Koch criteria: poor/slight (κ < 0.21), fair (κ = 0.21–0.40), moderate (κ = 0.41–0.60), substantial (κ = 0.61–0.80), and almost perfect (κ = 0.81–1.00).

Panel A. Category redistribution (n = 600 reader-case observations)

Category	Without ancillary features, n (%)	With ancillary features, n (%)	Change, n
Equivocal	63 (10.5%)	35 5.8%)	- 28
Viable	57 (9.5%)	85 (14.2%)	+28

Panel B. Ancillary features present among reclassified observations (n = 28)

Feature	n	%
T2 hyperintensity positive	13	46.4%
Restricted diffusion positive	25	89.3%
Both features positive	10	35.7%
T2 positive only	3	10.7%
Restricted diffusion positive only	15	53.6%

Table 4. Effect of ancillary features on LR-TRA v2024 categorization. (A) Distribution of LR-TRA v2024 categories before and after application of ancillary features. (B) Ancillary features present among reclassified reader-case observations.

FIGURES



Figure 1. 76-year-old man with HCC in segments II/III. MRI with fat-suppressed contrast-enhanced T1-weighted arterial phase (A), T2 (B), and DWI (C). The lesion showed equivocal enhancement on the post-contrast sequence (arrow in A), moderate T2 hyperintensity (arrow in B), and restricted diffusion (arrow in C). Based on the post-contrast sequence alone, the lesion would be categorized as LR-TR Equivocal according to LR-TRA v2024. However, application of both ancillary features upgrades the lesion from LR-TR Equivocal to LR-TR Viable.

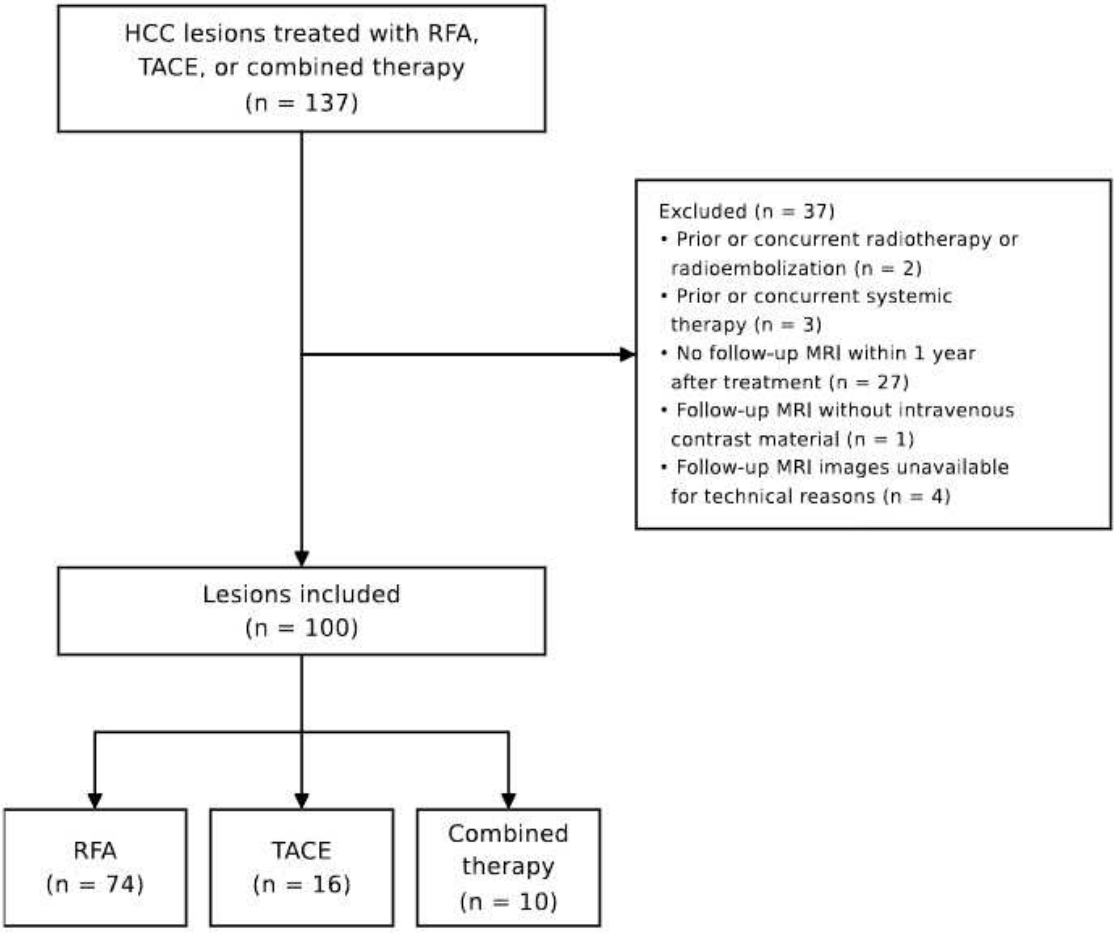


Figure 2. Flowchart of lesion selection based on inclusion and exclusion criteria

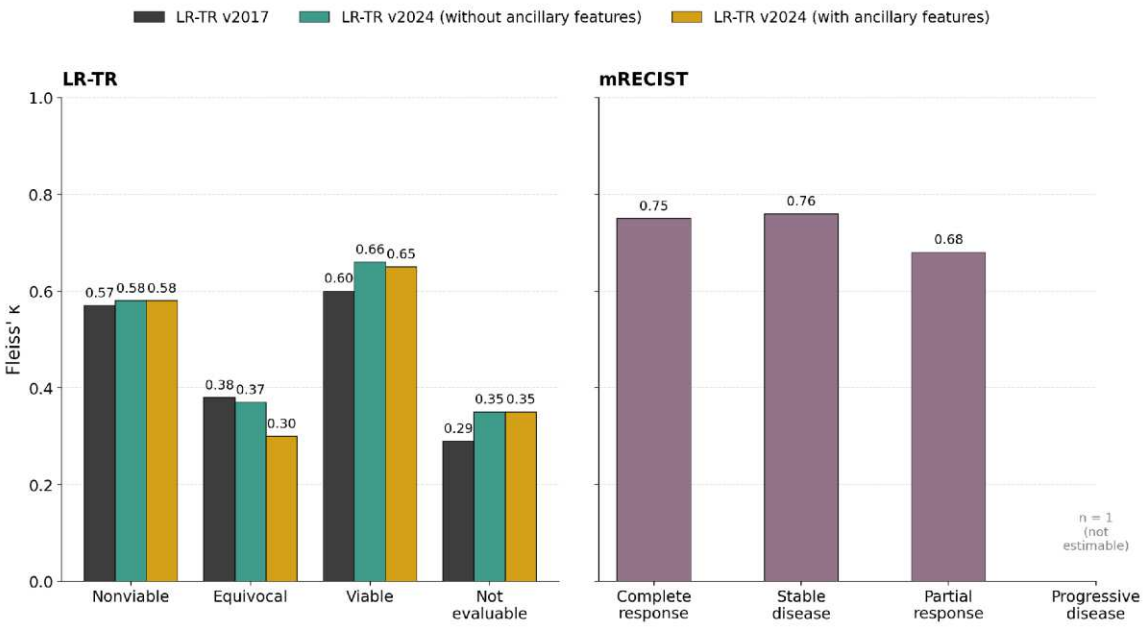


Figure 3. Fleiss' κ by category, for LR-TRA (v2017, v2024 without/with ancillary features) and mRECIST.

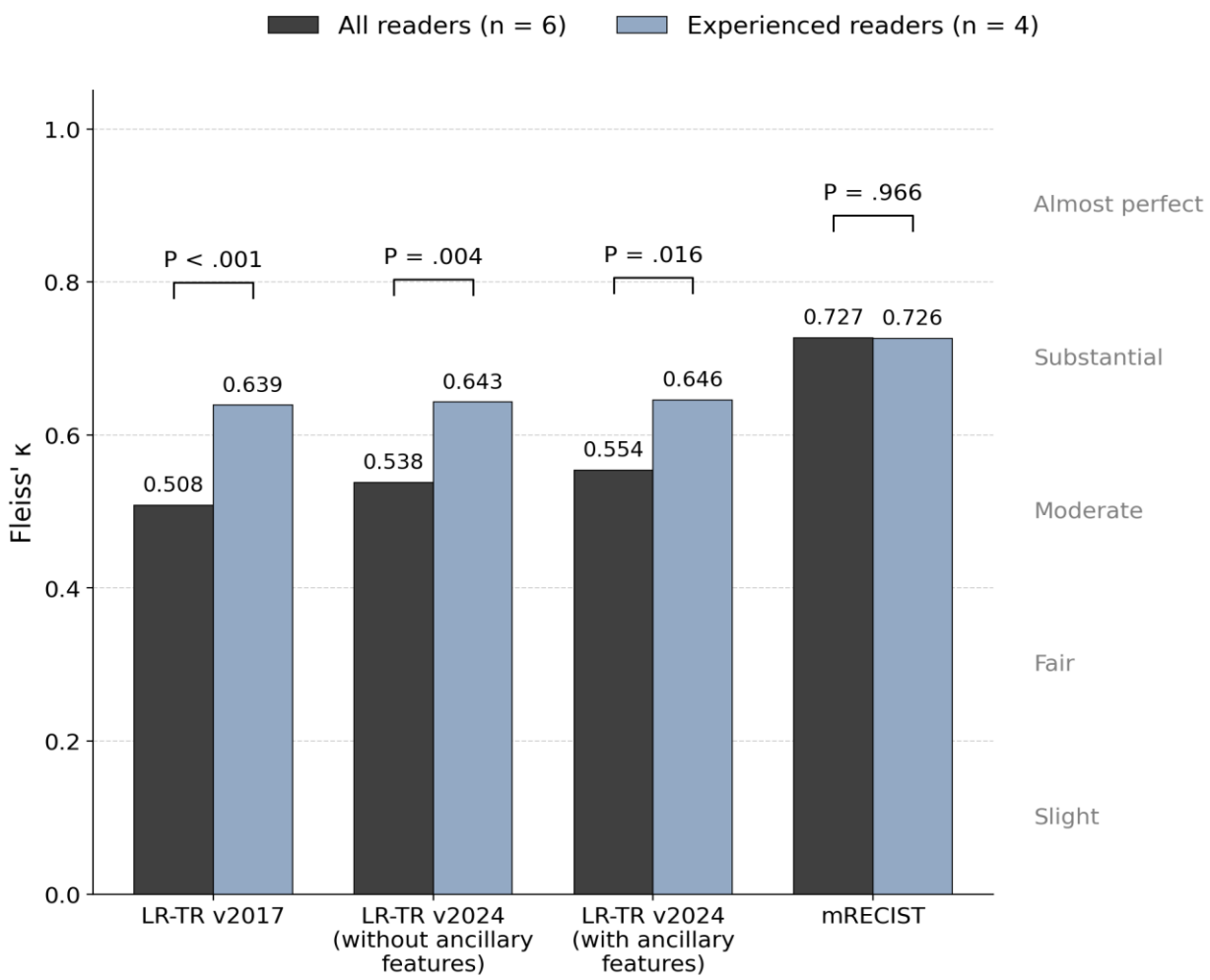


Figure 4. Global Fleiss' κ by classification system: all readers vs. subspecialized-reader subgroup.

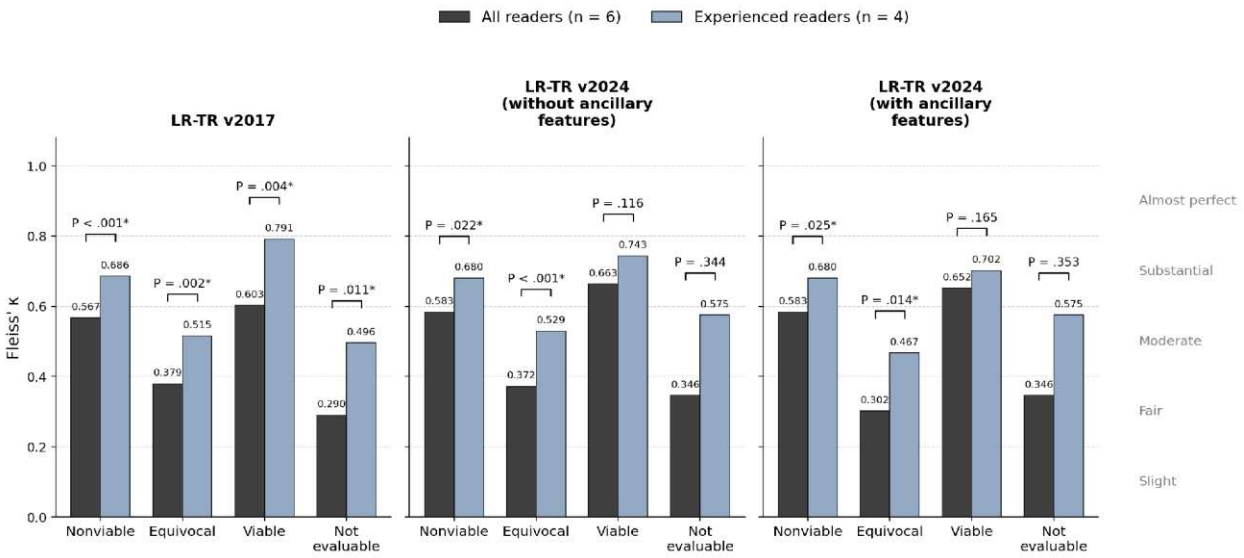


Figure 5. Category-specific Fleiss' κ, all readers (n = 6) vs. subspecialized readers (n = 4), with bootstrap significance testing, for LR-TRA v2017, v2024 without ancillary features, and v2024 with ancillary features.

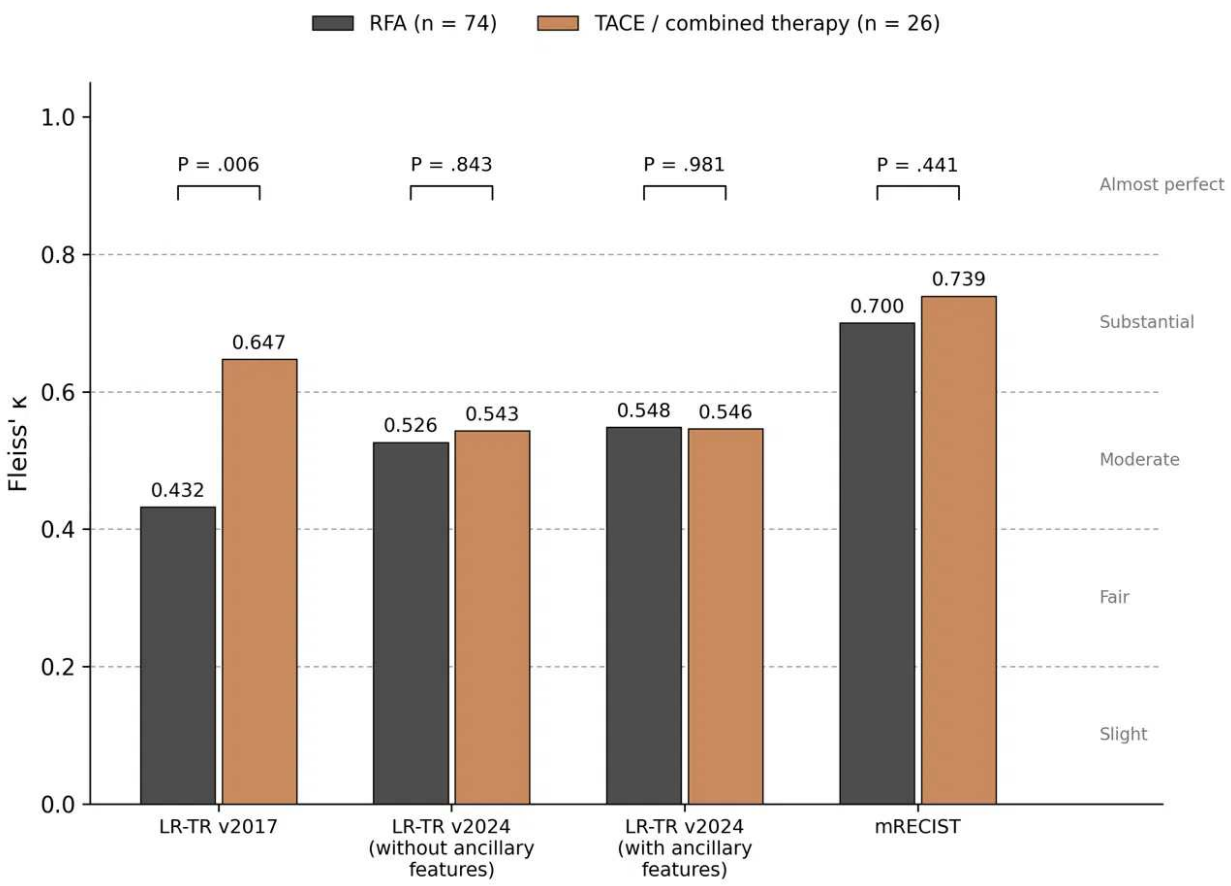


Figure 6. Global Fleiss' κ by classification system: RFA vs. TACE/combined therapy.

SUPPLEMENTARY MATERIAL

Classification system	Fleiss' κ, unweighted (6 readers)	Cohen's κ, linear-weighted (6 readers)	Cohen's κ, quadratic-weighted (6 readers), 95% CI	Cohen's κ, linear-weighted (4 subspecialized)	Cohen's κ, quadratic-weighted (4 subspecialized), 95% CI
LR-TRA v2017	0.508	0.631	0.693 (0.625–0.761)	0.743	0.808 (0.755–0.861)
LR-TRA v2024, without ancillary features	0.538	0.634	0.702 (0.651–0.753)	0.717	0.777 (0.732–0.823)
LR-TRA v2024, with ancillary features	0.554	0.631	0.670 (0.626–0.714)	0.704	0.737 (0.692–0.781)

Table S1. Sensitivity analysis comparing unweighted Fleiss' κ with linear-weighted and quadratic-weighted Cohen's κ by LR-TRA classification system and reader group. Agreement was interpreted according to conventional Landis–Koch criteria: poor/slight (κ < 0.21), fair (κ = 0.21–0.40), moderate (κ = 0.41–0.60), substantial (κ = 0.61–0.80), and almost perfect (κ = 0.81–1.00). *95% CIs for the quadratic-weighted estimates were derived from the standard deviation of the pairwise reader-pair κ values (15 pairs for the six-reader cohort and 6 pairs for the four-reader subgroup) and should therefore be interpreted as descriptive of the variability across reader pairs rather than as formal large-sample variance estimates.*

Study	Imaging	n patients	Readers (experience)	Treatment	Classification	Kappa / Overall agreement
Abdel Razek et al. 2020 (7)	MRI	93	2 experts (25, 10 y)	TACE + thermal ablation; prospective	LR-TRA v2017	Weighted, **κ = 0.938**
Bartnik et al. 2022 (6)	CT	102	2 (8, 5 y)	TACE	LR-TRA v2017/2018	Unweighted, Cohen κ, **κ = 0.70**
Zhou et al. 2025 (11)	MRI	231	2 (5, 10 y)	TACE ± ablation	LR-TRA v2017/v2024; mRECIST	Weighted, **κ = 0.72–0.75** (LR-TR); κ = 0.67 (mRECIST)
Shin et al. 2026 (9)	MRI (hepatobiliary agent); CT subgroup	245	3 (13, 9, 7 y)	TACE, ablation, or combined; surgical pathology	LR-TRA v2017/v2024 ± AF	Unweighted, Fleiss κ, κ = 0.52 (v2017) and 0.54 (v2024) at MRI; 0.65 at CT
Wang et al. 2024 (4)	MRI	99	3 (10, 6, 2 y)	TACE + RFA	LR-TRA v2017	Unweighted, Fleiss κ, **κ = 0.704**
Castiglione et al. 2025 (20)	MRI	78	2 (5, 9 y)	MWA vs. TAE	LR-TRA v2017	Unweighted, Cohen κ, **κ = 0.60**; 0.65 MWA vs. 0.51 TAE
Kim SW et al. 2020 (17)	MRI	183	2	TACE + RFA	LR-TRA v2017	Weighted κ, **κ = 0.58**
Kim YY et al. 2022 (19)	MRI	141	2 (~1 y) + adjudicator (32 y)	TACE + ablation + RT	LR-TRA v2017 + individual AF	Unweighted, Cohen κ, **κ = 0.73–0.97** by feature
Cools et al. 2020 (21)	MRI	45	2 experts (>9 y)	Thermal ablation; bridge to transplant	LR-TRA v2017	Unweighted, Cohen κ, **κ = 0.75**

Study	Imaging	n patients	Readers (experience)	Treatment	Classification	Kappa / Overall agreement
Kim DW et al. 2021 (18)	CT + MRI	851 (pooled, 8 studies)	Mostly 2; heterogeneous experience	Ablation + TACE	LR-TRA v2017	Mixed, **pooled κ = 0.70** (95% CI, 0.58–0.82)
Present study, 2026	MRI	97	6; 4 subspecialists, 1 general radiologist, 1 resident	RFA + TACE + combined	LR-TRA v2017; v2024 ± AF; mRECIST	Unweighted, Fleiss κ, **κ = 0.508, 0.538, 0.554 (LR-TRA); 0.727 (mRECIST)**; 0.64–0.65 among subspecialists

Table S2. Comparison of published studies evaluating interobserver agreement for post-treatment HCC assessment using LR-TRA and related classification systems. Studies are summarized according to imaging modality, reader experience, treatment type, classification system, kappa methodology, and overall interobserver agreement. The LR-TRA was unchanged between v2017 and v2018. AF = ancillary features; CI = confidence interval; CT = computed tomography; HCC = hepatocellular carcinoma; LR-TRA = LI-RADS treatment response algorithm; mRECIST = modified Response Evaluation Criteria in Solid Tumors; MRI = magnetic resonance imaging; MWA = microwave ablation; RFA = radiofrequency ablation; RT = radiotherapy; TACE = transarterial chemoembolization; TAE = transarterial embolization; y = years.

BIBLIOGRAPHY

1. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin. 2021 May;71(3):209–49. doi:10.3322/caac.21660

2. Aslam A, Chernyak V, Miller FH, Bashir M, Do R, Sirlin C, et al. CT/MRI LI-RADS 2024 Update: Treatment Response Assessment. Radiology. 2024 Nov 1;313(2):e232408. doi:10.1148/radiol.232408

3. Bae JS, Lee JM, Yoon JH, Kang H jin, Jeon SK, Joo I, et al. Evaluation of LI-RADS Version 2018 Treatment Response Algorithm for Hepatocellular Carcinoma in Liver Transplant Candidates: Intraindividual Comparison between CT and Hepatobiliary Agent–enhanced MRI. Radiology. 2021 May;299(2):336–45. doi:10.1148/radiol.2021203537

4. Wang Y, Asayo H, Wang W, Xu H, Yang D, Xu L, et al. Inter-reader agreement of LI-RADS treatment response algorithm among three readers with different seniorities for hepatocellular carcinoma after locoregional therapy. Acta Radiol. 2024 Dec;65(12):1458–64. doi:10.1177/02841851241289130

5. Shropshire EL, Chaudhry M, Miller CM, Allen BC, Bozdogan E, Cardona DM, et al. LI-RADS Treatment Response Algorithm: Performance and Diagnostic Accuracy. Radiology. 2019 Jul;292(1):226–34. doi:10.1148/radiol.2019182135

6. Bartnik K, Podgórska J, Rosiak G, Korzeniowski K, Rowiński O. Inter-observer agreement using the LI-RADS version 2018 CT treatment response algorithm in patients with hepatocellular carcinoma treated with conventional transarterial chemoembolization. Abdom Radiol. 2022 Jan;47(1):115–22. doi:10.1007/s00261-021-03272-9

7. Abdel Razek AAK, El-Serougy LG, Saleh GA, Shabana W, Abd El-Wahab R. Reproducibility of LI-RADS treatment response algorithm for hepatocellular carcinoma after locoregional therapy. Diagn Interv Imaging. 2020 Sep;101(9):547–53. doi:10.1016/j.diii.2020.03.008

8. Gupta P, Bansal A, Das GC, Kumar-M P, Chaluvashetty SB, Bhujade H, et al. Diagnostic accuracy of Liver Imaging Reporting and Data System locoregional treatment response criteria: a systematic review and meta-analysis. Eur Radiol. 2021 Oct;31(10):7725–33. doi:10.1007/s00330-021-07837-6

9. Shin J, Yoo J, Gu K, Baek SY, Bae JS, Lee MW, et al. Evaluating Accuracy of LI-RADS Nonradiation Treatment Response Algorithm v2024 and Ancillary Features at Hepatobiliary MRI versus CT. Radiology. 2026 Jun 1;319(3):e253143. doi:10.1148/radiol.253143

10. Zare M, Mohebbi A, Abdi A, Mohammadi A. Performance of the LI-RADS treatment response algorithm v2024: a systematic review and meta-analysis of response assessment accuracy and non-definitive categories. Abdom Radiol. 2026 Feb 6;51(8):3870–82. doi:10.1007/s00261-026-05380-w

11. Zhou S, Zhou G, Shen Y, Xia T, Zhao B, Sun Z, et al. LI-RADS Nonradiation Treatment Response Algorithm Version 2024: Diagnostic Performance and Impact of Ancillary Features. Am J Roentgenol. 2025 Feb;224(2):e2432035. doi:10.2214/AJR.24.32035

12. Rosenkrantz AB, Ginocchio LA, Cornfeld D, Froemming AT, Gupta RT, Turkbey B, et al. Interobserver Reproducibility of the PI-RADS Version 2 Lexicon: A Multicenter Study of Six Experienced Prostate Radiologists. Radiology. 2016 Sep;280(3):793–804. doi:10.1148/radiol.2016152542

13. Mussi TC, Yamauchi FI, Tridente CF, Tachibana A, Tonso VM, Recchimuzzi DZ, et al. Interobserver agreement of PI-RADS v.2 lexicon among radiologists with different levels of experience. J Magn Reson Imaging. 2020 Feb;51(2):593–602. doi:10.1002/jmri.26882

14. Harris PA, Taylor R, Thielke R, Payne J, Gonzalez N, Conde JG. Research electronic data capture (REDCap)—A metadata-driven methodology and workflow process for providing translational research informatics support. J Biomed Inform. 2009 Apr;42(2):377–81. doi:10.1016/j.jbi.2008.08.010

15. Landis JR, Koch GG. The Measurement of Observer Agreement for Categorical Data. Biometrics. 1977 Mar;33(1):159–74. doi:10.2307/2529310

16. Fleiss JL, Levin B, Paik MC. Statistical Methods for Rates and Proportions. 3rd ed. Hoboken, NJ: John Wiley & Sons; 2003.

17. Kim SW, Joo I, Kim HC, Ahn SJ, Kang HJ, Jeon SK, et al. LI-RADS treatment response categorization on gadoxetic acid-enhanced MRI: diagnostic performance compared to mRECIST and added value of ancillary features. Eur Radiol. 2020 May;30(5):2861–70. doi:10.1007/s00330-019-06623-9

18. Kim DW, Choi SH, Lee JS, Kim SY, Lee SJ, Byun JH. Interreader Reliability of Liver Imaging Reporting and Data System Treatment Response: A Systematic Review and Meta-Analysis. Diagnostics. 2021 Feb 4;11(2):237. doi:10.3390/diagnostics11020237

19. Kim YY, Kim MJ, Yoon JK, Shin J, Roh YH. Incorporation of Ancillary MRI Features Into the LI-RADS Treatment Response Algorithm: Impact on Diagnostic Performance After Locoregional Treatment of Hepatocellular Carcinoma. Am J Roentgenol. 2022 Mar;218(3):484–93. doi:10.2214/AJR.21.26677

20. Castiglione DG, Porreca A, Falsaperla D, Libra F, David E, Maiuzzo R, et al. Inter-Reader Agreement in LR-TRA Application and NLR Association in HCC Patients Treated with Endovascular vs. Ablative Procedures. Cancers. 2025 Feb 1;17(3):492. doi:10.3390/cancers17030492

21. Cools KS, Moon AM, Burke LMB, McGinty KA, Strassle PD, Gerber DA. Validation of the Liver Imaging Reporting and Data System Treatment Response Criteria After Thermal Ablation for Hepatocellular Carcinoma. Liver Transpl. 2020 Feb;26(2):203–14. doi:10.1002/lt.25673

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SUPPLEMENTARYMATERIAL.docx