Supplementary Material 4. Statistical Analysis

Sample Size Estimation
For sample size estimation, the expected per-lesion sensitivity of dynamic contrast-enhanced CT and MRI for HCC were assumed to be 68% and 80%, respectively, with the same specificity of 94% (1). Based on these expected values, we calculated that we would need 209 or more hepatic lesions to obtain a statistical power of 80% to detect a significant difference in diagnostic performance with a two-sided type I error of 5% (2).

Frequencies of Features
The baseline characteristics were compared using ANOVA for categorical variables and chi-square test for categorical variables. Multiple comparison was corrected using Bonferroni’s method. We used Fisher’s exact test or chi-square test to compare the frequency of major features.

Diagnostic Performance
We used a generalized estimating equation (GEE) method to calculate and compare the diagnostic performance for 1) diagnosing hepatic malignancies (i.e., HCC plus other malignancies vs. benign) with LR-5, -5V, or -M considered positive for the diagnosis, 2) differentiating HCC from non-HCC lesions (i.e., HCC vs. other malignancies plus benign) with LR-5 or -5V considered positive for HCC, and 3) differentiating non-HCC malignancies from other lesions (i.e., other malignancies vs. HCC plus benign) with LR-M considered positive for other malignancies. Next, the diagnostic performance was compared in subgroups only containing two disease entities using the GEE method: for 1) differentiating HCC from benign lesions (with other malignancies excluded from the analysis) and 2) differentiating between HCC and other malignancies (with benign lesions excluded). In addition, we examined the diagnostic performance of LR-5 (without LR-5V) for HCC. Lastly, in order to examine how the diagnostic performance varies according to the dynamic phase used for evaluating washout appearance at gadoxetate-enhanced MRI, we examined the diagnostic performance using the portal phase alone for washout evaluation and compared it with the results obtained using both the portal and transitional phases. The GEE analysis was conducted using SAS 9.2 (SAS Institute Inc., Cary, NC, USA), with two-sided p values < 0.05 considered as indicating statistical significance.

Reference Standard
Pathologic diagnosis was used as a reference standard. Experienced liver pathologists in each participating site examined the specimens and made pathologic diagnoses based on histologic features and immunohistochemical findings, according to the World Health Organization guidelines (3). Combined hepatocellular and cholangiocarcinoma (or biphenotypic hepatic carcinoma) was considered non-HCC malignancies. During pathologic examination, the pathologists had access to clinical history and imaging reports, but were unaware of the results of LI-RADS evaluation which was performed after the pathologic diagnosis.

Net Reclassification Improvement
To examine the added value of ancillary features, we tabulated the LI-RADS categories before and after applying ancillary features, and calculated a net reclassification improvement (NRI). NRI is a statistic to assess the improvement in prediction performance gained by adding a new factor to a model by measuring the extent to which individual subjects, with and without disease, are appropriately reclassified into more appropriate higher- or lower-risk categories (4). For our NRI calculation, cases categorized as LR-M as well as ‘not-visible’ cases were excluded, because LR-M does not follow the same ordinal likelihood as other categories and very few cases (0.4%, 4/918) were changed from or into LR-M after using ancillary features. This analysis was not performed on CT cases, because most of the ancillary features were evaluable only with MRI. NRI was calculated using the ‘reclassification’ function of R software (version 3.3.3, The R Foundation for Statistical Computing, Vienna, Austria).

Inter-Reader Agreement
We computed kappa statistics as indices of inter-reader agreement between four readers on the LI-RADS category and the presence or absence of major features, using the R function ‘Kappam.fleiss’. A kappa statistic of 0.89−1.00 was considered excellent agreement, 0.60−0.80 was good agreement, 0.40−0.59 was moderate agreement, 0.20−0.39 was fair agreement, and 0.00−0.19 was poor agreement.

REFERENCES