Hepatology International (2022) 16:1075–1084
https://doi.org/10.1007/s12072-022-10373-0

ORIGINAL ARTICLE

CT-based visual grading system for assessment of hepatic steatosis: diagnostic performance and interobserver agreement

Hyo Jung Park[1] · Kyoung Won Kim[1] · Heon-Ju Kwon[2] · Sunyoung Lee[3] · Dong Wook Kim[1] · Hye Hyeon Moon[1] · Gi-Won Song[4] · Sung-Gyu Lee[4]

Received: 24 March 2022 / Accepted: 30 May 2022 / Published online: 5 July 2022

Abstract

Background Hepatic steatosis (HS) can be comprehensively assessed by visually comparing the hepatic and vessel attenuation on unenhanced computed tomography (CT). We aimed to evaluate the reliability and reproducibility of a CT-based visual grading system (VGS) for comprehensive assessment of HS.

Methods In this retrospective study, a four-point VGS based on the visual comparison of liver and hepatic vessels was validated by six reviewers with diverse clinical experience using the unenhanced CT images of 717 potential liver donors. The diagnostic performance of VGS and quantitative indices (difference and ratio of the hepatic and splenic attenuation) to diagnose HS were evaluated using multi-reader multi-case receiver operating characteristics (ROC) analysis (reference: pathology). The interobserver agreement was assessed using Fleiss κ statistics.

Results Using the VGS, all six reviewers showed areas under the ROC curves (AUROCs) higher than 0.9 for diagnosing total steatosis (TS) ≥ 30%, macrovesicular steatosis (MaS) ≥ 30%, and MaS ≥ 10%. No difference was noted between the AUROCs of the VGS and quantitative indices ($p \geq 0.1$). The reviewers showed substantial agreement (Fleiss κ, 0.61). Most discrepancies occurred between the two lowest grades of VGS (81.5%; 233/283), in which most subjects (97.0%; 226/233) had a MaS < 10%. The average-reader sensitivity and specificity of the VGS were 0.80 and 0.94 to detect TS ≥ 30% and 0.93 and 0.81 to detect MaS ≥ 10%.

Conclusion VGS was reliable and reproducible in assessing HS. It may be useful as a non-invasive and simple tool for comprehensive HS assessment.

✉ Kyoung Won Kim
kimkw@amc.seoul.kr

Hyo Jung Park
happyeahj@gmail.com

Heon-Ju Kwon
umkhj@hanmail.net

Sunyoung Lee
carnival0126@gmail.com

Dong Wook Kim
fsnoruen@gmail.com

Hye Hyeon Moon
aslhey@hanmail.net

Gi-Won Song
drsong71@amc.seoul.kr

Sung-Gyu Lee
sglee2@amc.seoul.kr

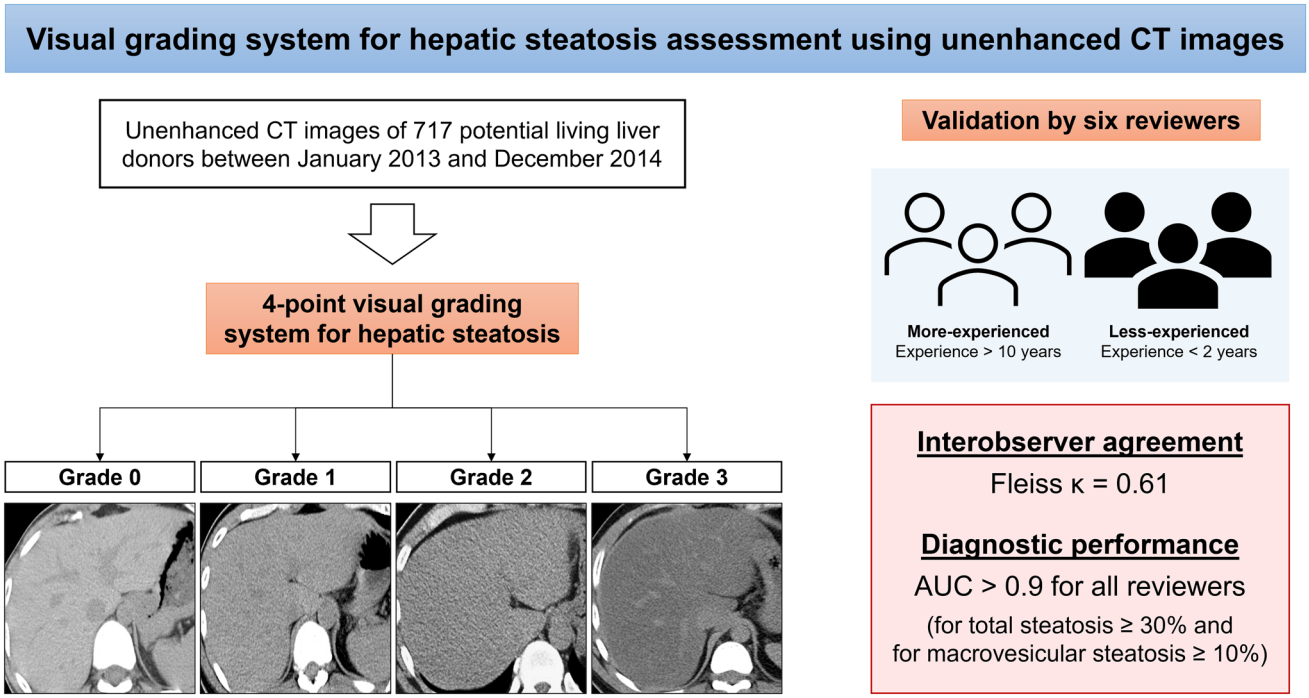
[1] Department of Radiology and Research Institute of Radiology, Asan Medical Center, University of Ulsan College of Medicine, 88 Olympic-Ro 43-Gil, Songpa-Gu, Seoul 05505, Republic of Korea

[2] Department of Radiology, Kangbuk Samsung Hospital, Sungkyunkwan University School of Medicine, Seoul, Republic of Korea

[3] Department of Radiology and Research Institute of Radiological Science, Severance Hospital, Yonsei University College of Medicine, Seoul, Republic of Korea

[4] Division of Liver Transplantation and Hepatobiliary Surgery, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea

Graphical abstract



Keywords Fatty liver · Computed tomography · Hepatic vessels · Visual score · Interobserver variability · Liver attenuation · Qualitative assessment · Liver donor · Transplantation · Diagnostic performance

Introduction

Hepatic steatosis (HS), the most prevalent liver disease worldwide [1], is a major factor in selecting appropriate donors for living donor liver transplantation [2, 3]. HS increases the risk of graft failure; while macrovesicular steatosis (MaS) greater than 10% is a concern, MaS greater than 30% is an absolute contraindication for donation [4]. In our institution, donors with a MaS of 10–15% and total steatosis (TS, sum of macro- and microvesicular steatosis) of 30% are preferred for right hemiliver donation, but donors with a moderate HS (30–50%) are allowed if they meet all of the following conditions: (1) age ≤ 35 years, (2) estimated ratio of the remnant left hemiliver to total liver volume ≥ 35%, (3) degree of the HS of the left hemiliver less than that of the right hemiliver, (4) sufficient graft-to-recipient weight ratio in the recipient, and (5) recipients unable to wait for weight reduction of the living donor candidates because of medical conditions requiring urgent liver transplantation. Living donor candidates not fulfilling the above criteria are reevaluated after weight reduction. Therefore, in potential liver donors with HS, repetitive assessment of liver fat content is required to monitor the degree of HS during the course of weight reduction.

Although liver biopsy is the reference standard for the HS assessment, its suitability for repetitive examinations is limited because of its invasiveness, low reproducibility, and high cost [5, 6]. Also, as the distribution of fat within the liver is spatially heterogenous [7–9], HS assessment using liver biopsy may be erroneous as only small pieces of hepatic tissue are obtained. Recent MRI techniques such as proton density fat fraction (PDFF) are considered accurate for assessing HS [10], but the high cost and limited accessibility of MRI may hamper its widespread use.

CT is widely used for the preoperative evaluation of potential liver donors. Using CT, intrahepatic fat can be quantitatively evaluated by placing a region-of-interest (ROI) and measuring the attenuation values (Hounsfield units; HU) of the liver relative to an internal reference, most commonly the spleen. Although there has been controversy over the performance of CT for HS assessment especially for mild HS [11], recent studies have shown that quantitative CT analysis can accurately assess mild HS and show good agreement with MRI PDFF and pathology [12, 13]. However, such a quantitative method is time-consuming and operator-dependent. Special care is required to reduce measurement error because the sampled attenuation values may be contaminated by environmental density artifact, partial

volume averaging, and other reactions that affect the attenuation of the target organs.

The degree of HS can be comprehensively assessed by visually comparing the attenuation of the liver and hepatic vessels, which is also non-invasive and simple to perform. The purpose of this study was to validate the reliability and reproducibility of a visual grading system (VGS) for diagnosing HS in a large cohort of living liver donor candidates.

Materials and methods

Study population

Living liver donor candidates who underwent abdominal CT and liver biopsy for pre-donation workup between January 2013 and December 2014 were consecutively identified. Inclusion criteria were: (a) subjects with abdominal CT scans; (b) at least two biopsy specimens; and (c) less than 3 months interval between CT and liver biopsy. Among 758 subjects initially recruited, 41 were excluded (24 with biopsy–CT interval of > 3 months; 17 without unenhanced CT images). Subjects who fulfilled the inclusion criteria were eligible, regardless of the degree of their HS. The remaining 717 subjects [480 men; median age, 28 years; interquartile range (IQR), 22–35] were finally included. This retrospective study was approved by the institutional review board of our institution and the requirement for written informed consent was waived.

CT protocol

CT scans were performed using a 64 multidetector scanner (Definition, Siemens, Erlangen, Germany). Unenhanced axial images were obtained during a single breath-hold followed by contrast-enhanced images with intravenous administration of 150 mL of iopromide (Ultravist 370; Bayer Schering Pharma, Berlin, Germany). The parameters for the unenhanced CT scans were beam collimation of 64 × 0.6 mm; spiral pitch of 1; gantry rotation time of 0.5 s; tube voltage of 100 kVp; and tube current of 120–200 mAs with automatic exposure control (Care Dose 4D; Siemens). Images were reconstructed with a section thickness of 5 mm at 5 mm intervals. The median interval between CT and biopsy was 9 days (IQR 5–17), and the interval was < 1 month in 87.9% of subjects (648/737).

Hepatic attenuation assessment

The VGS for HS assessment was validated by six reviewers, including three more-experienced radiologists (with 20, 13, and 10 years of experience in liver imaging, respectively) and three less-experienced radiologists (two fellows with 1–2 years of training experience in liver imaging and a 3rd-year radiology resident). All reviewers were blinded to the subjects' medical histories, previous imaging, and pathologic reports.

A four-point VGS based on comparisons of the brightness of the liver with those of hepatic vessels on unenhanced CT images was developed as follows: grade 0 (G0; hepatic vessels showing lower attenuation than the hepatic parenchyma, with no or minimal margin blurring in less than one-third of the liver), grade 1 (G1; hepatic vessels showing lower attenuation than hepatic parenchyma but with margin blurring in more than one-third of the liver), grade 2 (G2; hepatic vessels showing the same attenuation as hepatic parenchyma), and grade 3 (G3; hepatic vessels showing higher attenuation than hepatic parenchyma; Fig. 1).

Grading assessments were performed in two sessions. The first session served as a training session and consisted of 20 selected cases: six with no pathologic HS (TS < 5%); five with mild pathologic HS (TS 5–30%); five with moderate pathologic HS (TS 30–60%); and four with severe pathologic HS (TS ≥ 60%). These data were prepared by one author not involved in the imaging assessment. Through discussion, a consensus for each case was reached among

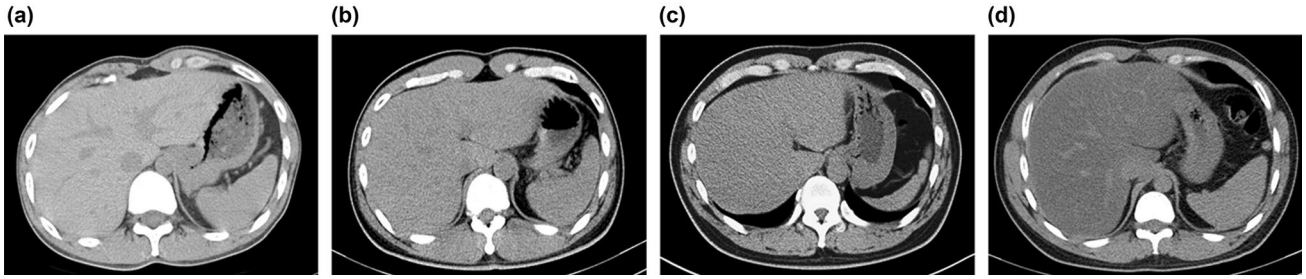


Fig. 1 Examples of each grade of the VGS for the assessment of hepatic steatosis. **a** G0, hepatic vessels showing lower attenuation than the hepatic parenchyma, with no or minimal margin blurring in less than one-third of the liver; **b** G1, hepatic vessels showing lower attenuation than the hepatic parenchyma, but with margin blurring in more than one-third of the liver; **c** G2, hepatic vessels showing the same attenuation as hepatic parenchyma; and **d** G3, hepatic vessels show higher attenuation than hepatic parenchyma. *VGS* visual grading system

the six reviewers. After a 2-week washout period, the main testing session was performed using images from all 717 subjects.

Two quantitative indices of liver attenuation were obtained from the unenhanced CT: $CT_{L\text{-}S}$ (mean hepatic attenuation minus mean splenic attenuation) and $CT_{L/S}$ (mean hepatic attenuation divided by mean splenic attenuation). Detailed information is provided in the Supplemental Materials.

Liver biopsy and histologic examination

All subjects underwent ultrasound-guided percutaneous liver biopsy by board-certified radiologists (all with > 4 years of experience). Liver tissue was obtained using an 18-gauge needle (TSK Stericut 18G coaxial; TSK Laboratory, Tochigi, Japan). Using the intercostal approach, two or three biopsy specimens, each approximately 1.5 cm in length, were obtained from two different sites in the right hepatic lobe. Specimens were stained with hematoxylin–eosin and Masson trichrome and were examined by board-certified pathologists with > 7 years of experience. TS was evaluated according to the percentage of hepatocytes containing fat granules. The degree of MaS was determined according to the percentage of hepatocytes in which a single large fat droplet or smaller fat droplets occupied the cytoplasm, pushing the nucleus to the periphery.

Statistical analysis

Correlations between imaging-assessed HS (VGS, $CT_{L\text{-}S}$, and $CT_{L/S}$) and pathologic HS were evaluated using Spearman's rank correlation coefficient or Kendall's rank correlation coefficient, as appropriate. The diagnostic performance of each method to detect clinically significant HS (i.e., TS ≥ 30%, MaS ≥ 30%, and MaS ≥ 10%) was evaluated by using receiver operating characteristics (ROC) curve analysis with pathology as the reference. Multi-reader multi-case (MRMC) ROC analysis was used to obtain the average of the six reviewers' areas under the ROC curves (AUROCs) and their 95% confidence intervals (CIs). AUROCs were compared using DeLong's method [14]. The performance of the VGS for distinguishing each stage of the TS (TS < 5%, TS 5–30%, TS 30–60%, and TS ≥ 60%) and MaS (MaS < 10%, MaS 10–30%, and MaS ≥ 30%) was evaluated by using the Obuchowski measure, a multinomial version of ROC analysis adopted for ordinal references [15]. Interobserver agreement among the six reviewers for the VGS was assessed using Fleiss κ statistics. Interobserver agreement was compared between the more- and less-experienced reviewer groups using z score and standard errors of the differences. Agreements between pairs of reviewers were assessed using weighted κ statistics. The sensitivity and specificity of the VGS for detecting clinically significant HS were calculated for each reviewer. The corresponding pooled estimates for average-reviewer were calculated using the generalized estimating equation logistic regression to adjust for clustering of the six-reviewer data of each patient. To obtain the sensitivity and specificity of $CT_{L\text{-}S}$ and $CT_{L/S}$, the optimal cutoffs were determined using Youden's index [16]. Statistical analyses were performed using R version 4.1.1 (R Foundation for Statistical Computing, Vienna, Austria). Two-sided p values < 0.05 were considered significant.

Results

Study population

The characteristics of the 717 subjects are summarized in Table 1. There were 457 (63.7%) subjects with TS < 5%, 239 (33.3%) with TS 5–30%, 17 (2.4%) with TS 30–60%, and four (0.6%) with TS ≥ 60%. For MaS, 634 subjects (88.4%) had MaS < 10%, 66 (9.2%) had MaS 10–30%, 30%, and 17 (2.4%) had MaS ≥ 30%. The median differences in

Table 1 Characteristics of the study population

Characteristic	Value
Number	717
Age (years)	28 (22–35)
Sex	
Men	480 (66.9)
Women	237 (33.1)
Height (m)	1.7 (1.6–1.8)
Body weight (kg)	65.7 (58.5–73.6)
BMI (kg/m^2)	22.8 (20.8–24.7)
Histologic fat content	
Total steatosis (%)	2.0 (0.0–5.0)
< 5%	457 (63.7)
5–30%	239 (33.3)
30–60%	17 (2.4)
≥ 60%	4 (0.6)
Macrovesicular steatosis (%)	1 (0.0–3.0)
< 10%	634 (88.4)
10–30%	66 (9.2)
≥ 30%	17 (2.4)
Laboratory data	
AST (IU/L)	18 (15–20)
ALT (IU/L)	14 (11–20)
Total bilirubin (IU/L)	0.6 (0.5–0.9)

Data are numbers with percentages in parentheses or median with interquartile range in parentheses

ALT alanine aminotransferase, *AST* aspartate aminotransferase, *BMI* body mass index

body weight and body mass index between the time of liver biopsy and CT examination were 0.9 kg (IQR, 0.4–1.9 kg) and 0.3 kg/m^2 (IQR, 0.1–0.6 kg/m^2), respectively, in all study subjects, and 2.1 kg (0.8–5.2 kg) and 0.7 kg/m^2 (IQR, 0.3–1.7 kg/m^2) in the 69 subjects with a biopsy–CT interval of more than one month.

Correlation between HS assessed by imaging and pathologic HS

The distribution of TS and MaS in each visual grade assigned by each reviewer is shown in Fig. 2. The VGS and pathologic HS (both TS and MaS) showed positive correlations for all six reviewers ($p < 0.001$; Table E1). For the six reviewers, the mean correlation coefficients between VGS and pathologic HS were 0.543 (range 0.494–0.572; $p < 0.001$) for TS and 0.6 (range 0.518–0.697; $p < 0.001$) for MaS. The median values of $CT_{L\text{-}S}$ and $CT_{L/S}$ were 9.7 (IQR 5.6–13.5) and 1.2 (IQR 1.1–1.3), respectively. $CT_{L\text{-}S}$ and $CT_{L/S}$ showed negative but slightly less correlation with pathologic HS compared to VGS (mean correlation coefficient for TS, − 0.311 for $CT_{L\text{-}S}$ and − 0.301 for $CT_{L/S;}$ $p < 0.001$; mean correlation coefficient for MaS, − 0.451 for $CT_{L\text{-}S}$ and − 0.366 for $CT_{L/S}$; $p < 0.001$).

Diagnostic performance of the imaging-based method to detect clinically significant HS

All six reviewers showed AUROCs > 0.9 for diagnosing TS ≥ 30%, MaS ≥ 30%, and MaS ≥ 10% (Table E2). The

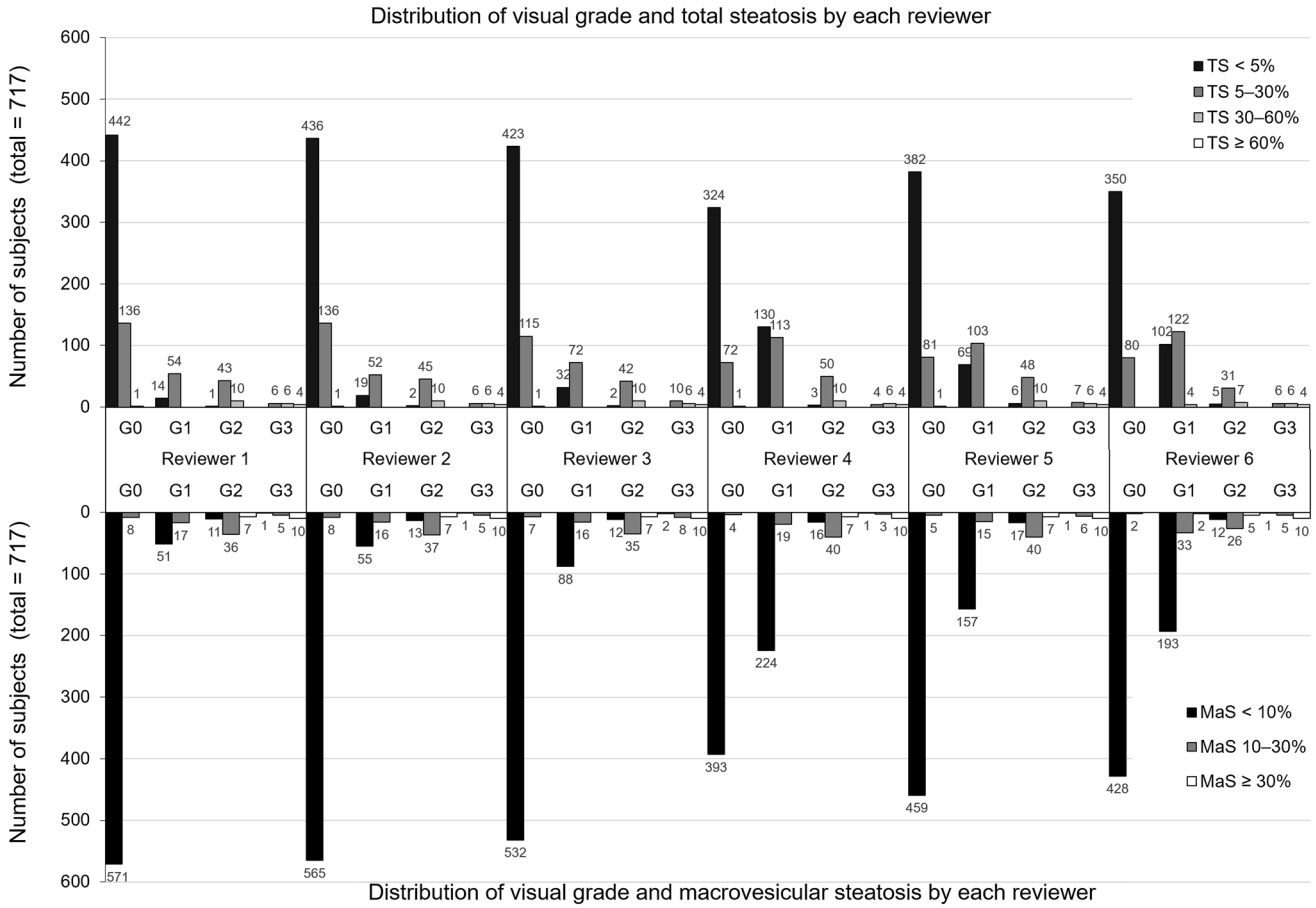


Fig. 2 Distribution of the visual grade assessments of each reviewer according to total steatosis and macrovesicular steatosis. Cases graded with G0 had TS < 5% (76.1–82.3%), TS 5–30% (17.5–23.7%), and TS ≥ 30% (0–0.3%); those graded as G1 had TS < 5% (20.6–53.5%), TS 5–30% (46.5–79.4%), and TS ≥ 30% (0–1.8%); those graded as G2 had TS < 5% (1.9–11.6%), TS 5–30% (72.1–98.1%), and TS ≥ 30% (0–18.5%); and those graded as G3 had TS < 5% (28.6–50%), TS 5–30% (30.0–42.9%), or TS ≥ 30% (20.0–28.6%). Regarding macrovesicular steatosis, those graded as G0 had MaS < 10% (98.6%–99.5%) and MaS 10–30% (0.5–1.4%); those graded as G1 had MaS < 10% (75.0–92.2%), MaS 10–30% (7.8–25%), and MaS ≥ 30% (0–0.9%); those graded as G2 had MaS < 10% (20.4–27.9%), MaS 10–30% (60.5–66.7%), and MaS ≥ 30% (10.9–13%); and those graded as G3 had MaS < 10% (5.9–10.0%), MaS 10–30% (21.4–40%), and MaS ≥ 30% (50.0–71.4%). Nearly all G0 and G1 cases had TS < 30% (99.4–99.8%) and MaS < 10% (94.4–96.9%). The majority of G2 and G3 cases had MaS ≥ 10% (77.8–96.7%). The visual grades and pathologic hepatic steatosis (both total and macrovesicular steatosis) showed a positive correlation for all six reviewers ($p < 0.001$). *MaS* macrovesicular steatosis, *TS* total steatosis

ROC curves and AUROCs of VGS, $CT_{L\text{-}S}$, and $CT_{L/S}$ for diagnosing TS ≥ 30%, MaS ≥ 30%, and MaS ≥ 10% are shown in Fig. E1 and Table 2. In the MRMC analysis, the AUROCs of VGS were 0.921 (95% CI 0.906–0.937) for TS ≥ 30%, 0.975 (95% CI 0.968–0.983) for MaS ≥ 30%, and 0.921 (95% CI 0.911–0.931) for MaS ≥ 10%. The ranges of the AUROCs of the quantitative methods were 0.887–0.890 for TS ≥ 30%, 0.981–0.982 for MaS ≥ 30%, and 0.896–0.902 for MaS ≥ 10%. No differences were noted between the AUROCs of VGS and the quantitative methods ($p \geq 0.1$).

The per-reviewer and average-reviewer sensitivity and specificity are shown in Table E3. VGS G2 was the best cutoff (i.e., G0–1 vs. G2–3) for detecting TS ≥ 30% and MaS ≥ 30%, and VGS G1 was the best cutoff (i.e., G0 vs. G1–3) for detecting MaS ≥ 10%. When using the VGS G2 as the cutoff, the average-reviewer sensitivity and specificity for detecting TS ≥ 30% were 0.80 and 0.94, respectively, and those for detecting MaS ≥ 30% were 0.98 and 0.92, respectively. When using the VGS G1 as the cutoff, the average-reviewer sensitivity and specificity for detecting MaS ≥ 10% were 0.93 and 0.81, respectively. Optimal cutoffs were 4.02 (TS ≥ 30%), 1.33 (MaS ≥ 30%), and 3.33 (MaS ≥ 10%) for $CT_{L\text{-}S}$, and 1.05 (TS ≥ 30%), 0.95 (MaS ≥ 30%), and 1.07 (MaS ≥ 10%) for $CT_{L/S}$. The average-reviewer VGS showed equivalent sensitivity compared to $CT_{L\text{-}S}$ and $CT_{L/S}$ except it showed a higher sensitivity than both quantitative parameters in detecting MaS ≥ 10%. The specificity of the average-reviewer VGS was higher than both quantitative parameters in detecting TS ≥ 30% and was equivalent to both quantitative parameters in detecting MaS ≥ 30% and MaS ≥ 10%. Using G1 as a cutoff, all six reviewers showed very few false-negative results when using the VGS to detect TS ≥ 30% and MaS ≥ 10% (Table E4).

For the prediction of TS and MaS, all the reviewers showed a higher Obuchowski measure for predicting TS (range 0.73–0.78) and MaS (range 0.91–0.93) than the quantitative parameters (0.72 for $CT_{L\text{-}S}$ and 0.71 for $CT_{L/S}$ in predicting TS and 0.90 for both $CT_{L\text{-}S}$ and $CT_{L/S}$ in predicting MaS) (Table E5). The diagnostic performance of each method for detecting TS > 5% is shown in Table E6.

Interobserver agreement of the VGS

The interobserver agreement with consistency and discrepancies among the reviewers are summarized in Table 3. All reviewers showed a substantial agreement ($\kappa = 0.61$; 95% CI 0.60–0.63) when using the VGS. All measures of interobserver agreement between two paired reviewers were > 0.75, representing substantial or almost perfect agreement (Table E7). Consistent grading from all reviewers was noted in 60.1% (431/717). More-experienced reviewers showed a higher agreement than less-experienced reviewers (κ, 0.82 and 0.68, respectively, $p < 0.001$; Table 3). The more-experienced reviewers showed consistent grading for 90.2% (647/717), higher than that achieved by the less-experienced reviewers (74.1%, 531/717; $p < 0.001$). Discrepancies in the more-experienced group occurred mostly between G0 and G1 (82.8%, 58/70). In the less-experienced group, discrepancies also occurred most frequently between G0 and G1 (77.4%, 144/186), followed by G1 and G2 (21.0%, 39/186).

Pathology of consistent and discrepant visual gradings

The pathologic results of the consistent and discrepant gradings are listed in Table 4. All G0 and G1 cases that were consistently graded by all reviewers had TS < 30% (100%, 391/391) and mostly had MaS < 10% (98.0%, 383/391). The vast majority of cases graded as G2 or G3 by all reviewers had MaS ≥ 10% (92.5%, 37/40).

Discrepancies in visual grading most commonly occurred between the assignments of G0 and G1 (81.5%, 233/286); pathologically, the vast majority of these cases were categorized as TS < 30% (99.5%, 232/233) and MaS < 10% (97.0%, 226/233). In the more-experienced group, all of the discrepant cases between G0 and G1 had TS < 30% (100%, 58/58), and mostly had MaS < 10% (94.8%, 55/58). Similarly, in the less-experienced group, most of the discrepant cases between G0 and G1 were TS < 30% (99.3%, 143/144) and MaS < 10% (97.2%, 140/144).

Table 2 AUROCs of imaging-based methods for predicting clinically significant HS

Parameters	AUROC			*p* value	
	VGS[a]	$CT_{L\text{-}S}$	$CT_{L/S}$	VGS vs. $CT_{L\text{-}S}$	VGS vs. $CT_{L/S}$
To differentiate TS ≥ 30% vs. TS < 30%	0.921 (0.906–0.937)	0.890 (0.812–0.967)	0.887 (0.806–0.966)	0.24	0.20
To differentiate MaS ≥ 30% vs. MaS < 30%	0.975 (0.968–0.983)	0.981 (0.966–0.996)	0.982 (0.967–0.997)	0.39	0.34
To differentiate MaS ≥ 10% vs. MaS < 10%	0.921 (0.911–0.931)	0.902 (0.861–0.944)	0.896 (0.852–0.940)	0.17	0.10

Data in parentheses are 95% confidence intervals

AUROC area under the receiver operating characteristics curve, $CT_{L\text{-}S}$ mean hepatic attenuation minus mean splenic attenuation, $CT_{L/S}$ mean hepatic attenuation divided by mean splenic attenuation, *HS* hepatic steatosis, *MaS* macrovesicular steatosis, *TS* total steatosis

[a]Multi-reviewer multi-case analysis

Table 3 Interobserver agreement, consistency, and discrepancy in the visual grading of HS

Parameters	All reviewers ($n=6$)	More-experienced ($n=3$)	Less-experienced ($n=3$)	p value
Fleiss κ	0.61 (0.60–0.63)	0.82 (0.79–0.85)	0.68 (0.65–0.72)	0.001*
Reviewers with consistent grading, no (%)				0.001†
Six reviewers	431 (60.1)	–	–	
Five reviewers	128 (17.9)	–	–	
Four reviewers	108 (15.1)	–	–	
Three reviewers	50 (7.0)	647 (90.2)	531 (74.1)	
Two reviewers	–	70 (9.8)	186 (25.9)	
Consistency among reviewers, no (%)				0.001†
G0	353 (49.2)	528 (73.6)	355 (49.5)	
G1	38 (5.3)	54 (7.5)	128 (17.9)	
G2	26 (3.6)	49 (6.8)	34 (4.7)	
G3	14 (2.0)	16 (2.2)	14 (2.0)	
Discrepancy among reviewers, no (%)				0.85†
G0 or G1	233 (32.5)	58 (8.1)	144 (20.1)	
G1 or G2	45 (6.3)	8 (1.1)	39 (5.4)	
G2 or G3	6 (0.8)	4 (0.6)	3 (0.4)	
G0 or G1 or G2[a]	2 (0.3)	–	–	

Data are numbers with percentages or 95% confidence intervals in parentheses

Visual grading was evaluated using a 4-point scale: G0 (hepatic vessels show lower attenuation than the hepatic parenchyma, with no or minimal margin blurring in less than one-third of the liver), G1 (hepatic vessels show lower attenuation than the hepatic parenchyma but with margin blurring in more than one-third of the liver), G2 (hepatic vessels show the same attenuation as that of hepatic parenchyma), and G3 (hepatic vessels show higher attenuation than the hepatic parenchyma)

*More- versus less-experienced; calculated using z scores

†More- versus less-experienced; calculated by Fisher's exact test or chi-squared test

[a]G0, G1, and G2 were assigned by one or more of the reviewers, respectively, in the same case

HS, hepatic steatosis

Discrepant grading between G1 and G2 occurred for 15.7% (45/286) of all discrepant cases among all reviewers, 11.4% (8/70) in the more-experienced group, and 21.0% (39/186) in the less-experienced group. The subjects with discrepancies in grading between G1 and G2 most commonly had TS of 5–30% (82.2% [37/45] among all six reviewers, 87.5% [7/8] in the more-experienced group, and 79.5% [31/39] in the less-experienced group), and 57.8% of them had MaS ≥ 10% (26/45) among all six reviewers, 62.5% (5/8) in the more-experienced group, and 56.4% (22/39) in the less-experienced group.

Discrepant grading between G2 and G3 was rare, occurring in only 3.1% (6/286) of all discrepant cases for all reviewers, 5.7% (4/70) for more-experienced reviewers, and 1.6% (3/186) for less-experienced reviewers; all subjects had TS of 5–30% and the majority of them had MaS of 10–30% (83.3%, 5/6; 75%, 3/4; and 100%, 3/3, respectively). Discrepant grading among G0 vs. G1 vs. G2 (i.e., G0, G1, and G2 assigned by one or more reviewers in the same case) occurred in two cases (0.3%, 2/717) among all six reviewers, with both subjects categorized as TS < 5% and MaS < 10%.

Discussion

In this study, we demonstrated that the four-point VGS based on visual comparison of hepatic and vessel attenuation on unenhanced CT allows reliable and reproducible assessment of HS. The VGS showed similar performance to the quantitative methods ($CT_{L/S}$ and $CT_{L\text{-}S}$) in diagnosing clinically significant HS. When evaluated by six radiologists with varying degrees of experience, the interobserver agreement was substantial, and most discrepancies did not involve misclassification of clinically significant HS. Therefore, our VGS can be a useful tool for non-invasive, simple, and integrative assessment of HS.

HS assessment by measuring attenuation values within ROIs drawn in liver parenchyma is a quantitative method [11, 17, 18] while VGS is a qualitative method. However, the qualitative nature of the VGS does not necessarily indicate that it is less objective than ROI-based CT attenuation measurement. Assessment of the HU value by placing the ROI on images is also operator-dependent. Density measurements in several areas of the liver can become

Table 4 Pathologic results for each consistent and discrepant category of the visual grading system for HS

Visual grading	Total steatosis							Macrovesicular steatosis					
	Degree	All reviewers		More-experienced		Less-experienced		Degree	All reviewers		More-experienced		Less-experienced
Consistent													
G0	< 5%	297	(41.4)	417	(58.2)	298	(41.6)	< 10%	351	(49)	522	(72.8)	353	(49.2)
	5–30%	56	(7.8)	110	(15.3)	57	(7.9)	10–30%	2	(0.3)	6	(0.8)	2	(0.6)
	30–60%	0	(0.0)	1	(0.1)	0	(0.0)	≥ 30%	0	(0.0)	0	(0.0)	0	(0.0)
	≥ 60%	0	(0.0)	0	(0.0)	0	(0.0)							
G1	< 5%	8	(1.1)	11	(1.5)	52	(7.3)	< 10%	32	(4.5)	42	(5.9)	117	(16.3)
	5–30%	30	(4.2)	43	(6.0)	76	(10.6)	10–30%	6	(0.8)	12	(1.7)	11	(1.5)
	30–60%	0	(0.0)	0	(0.0)	0	(0.0)	≥ 30%	0	(0.0)	0	(0.0)	0	(0.0)
	≥ 60%	0	(0.0)	0	(0.0)	0	(0.0)							
G2	< 5%	0	(0.0)	1	(0.1)	2	(0.3)	< 10%	2	(0.3)	10	(1.4)	6	(0.8)
	5–30%	19	(2.6)	38	(5.3)	25	(3.5)	10–30%	19	(2.6)	32	(4.5)	23	(3.2)
	30–60%	7	(1.0)	10	(1.4)	7	(1)	≥ 30%	5	(0.7)	7	(1.0)	5	(0.7)
	≥ 60%	0	(0.0)	0	(0.0)	0	(0.0)							
G3	< 5%	0	(0.0)	0	(0.0)	0	(0.0)	< 10%	1	(0.1)	1	(0.1)	1	(0.1)
	5–30%	4	(0.6)	6	(0.8)	4	(0.6)	10–30%	3	(0.4)	5	(0.7)	3	(0.4)
	30–60%	6	(0.8)	6	(0.8)	6	(0.8)	≥ 30%	10	(1.4)	10	(1.4)	10	(1.4)
	≥ 60%	4	(28.6)	4	(0.6)	4	(0.6)							
Discrepant														
G0 vs. G1	< 5%	145	(20.2)	27	(3.8)	100	(13.9)	< 10%	226	(31.5)	55	(7.7)	140	(19.5)
	5–30%	87	(12.1)	31	(4.3)	43	(6.0)	10–30%	7	(1.0)	3	(0.4)	4	(0.6)
	30–60%	1	(0.1)	0	(0.0)	1	(0.1)	≥ 30%	0	(0.0)	0	(0.0)	0	(0.0)
	≥ 60%	0	(0.0)	0	(0.0)	0	(0.0)							
G1 vs. G2	< 5%	5	(0.7)	1	(0.1)	5	(0.7)	< 10%	19	(2.6)	3	(0.4)	17	(2.4)
	5–30%	37	(5.2)	7	(1.0)	31	(4.3)	10–30%	24	(3.3)	5	(0.7)	20	(2.8)
	30–60%	3	(0.4)	0	(0.0)	3	(0.4)	≥ 30%	2	(0.3)	0	(0.0)	2	(0.3)
	≥ 60%	0	(0.0)	0	(0.0)	0	(0.0)							
G2 vs. G3	< 5%	0	(0.0)	0	(0.0)	0	(0.0)	< 10%	1	(0.1)	1	(0.1)	0	(0.0)
	5–30%	6	(0.8)	4	(0.6)	3	(0.4)	10–30%	5	(0.7)	3	(0.3)	3	(0.4)
	30–60%	0	(0.0)	0	(0.0)	0	(0.0)	≥ 30%	0	(0.0)	0	(0.0)	0	(0.0)
	≥ 60%	0	(0.0)	0	(0.0)	0	(0.0)							
G0 vs. G1 vs. G2[a]	< 5%	2	(0.3)	0	(0.0)	0	(0.0)	< 10%	2	(0.3)	0	(0.0)	0	(0.0)
	5–30%	0	(0.0)	0	(0.0)	0	(0.0)	10–30%	0	(0.0)	0	(0.0)	0	(0.0)
	30–60%	0	(0.0)	0	(0.0)	0	(0.0)	≥ 30%	0	(0.0)	0	(0.0)	0	(0.0)
	≥ 60%	0	(0.0)	0	(0.0)	0	(0.0)							

Data are numbers with percentages in parentheses

Visual grading was evaluated using a 4-point scale: G0 (hepatic vessels show lower attenuation than the hepatic parenchyma, with no or minimal margin blurring in less than one-third of the liver), G1 (hepatic vessels show lower attenuation than the hepatic parenchyma but with margin blurring in more than one-third of the liver), G2 (hepatic vessels show the same attenuation as that of hepatic parenchyma), and G3 (hepatic vessels show higher attenuation than the hepatic parenchyma)

[a]In the same case, G0, G1, and G2 were assigned at least once by one or more reviewers

erroneous by image artifacts [19, 20] as very dense bony structures are located close to the dome and the posterior section of the right hepatic lobe, beam hardening artifact frequently occurs and results in decreased attenuation of the affected area [21], and heterogeneous high- and low-density areas are frequently present in the left lateral section of the liver because of the edge gradient effect caused by pronounced differences in density between liver parenchyma and air in the stomach [21]. Also, similar to the case in liver biopsy, ROI acquisition is limited by sampling bias. The distribution of fat content within the liver is spatially heterogeneous [7–9] which is difficult to comprehensively assess by using ROIs that only allow the assessment of the liver in confined areas. Although the sampling error

could be resolved by drawing many ROIs or by including the whole liver as an ROI, this would be laborious and time-consuming. Moreover, there is no consensus regarding the cutoff value of hepatic attenuation for determining the presence and degree of HS, which precludes the use of the attenuation value in clinical practice. The aforementioned problems could be largely resolved by visual inspection of the whole liver. By using the VGS, we could comprehensively evaluate the overall density of the whole hepatic parenchyma and could thus be free from the bias of ROI measurement.

For a diagnostic tool to be successfully implemented in clinical practice, it is essential to thoroughly validate its performance according to its purpose [22]. The tool should demonstrate acceptable reliability and reproducibility in capturing the clinically meaningful problems and should be validated in a sufficiently large number of the target population [23]. We demonstrated the reliability of our VGS in a large cohort ($n = 717$) by showing positive correlations between visual grades and pathologic HS of the liver ($p < 0.001$), and high AUROCs for assessing clinically significant HS (> 0.9). The performance of the VGS was similar to those of quantitative assessments, and nearly all cases of G0–G1 had a TS < 30% (99.4–99.8%) and a MaS < 10% (94.4–96.9%).

The main focus of this study was to demonstrate the reliability and reproducibility of the VGS as a simple and comprehensive method for HS assessment, rather than to show that the VGS is superior to the standard quantitative parameters. We consider that both the VGS and the quantitative parameters could be used in a complementary manner to improve the accuracy of CT-based HS assessment (for example, if the decision is difficult with the VGS alone, the reader could measure the HU of liver, or vice versa), and this should be further investigated in future studies.

We demonstrated the reproducibility of the VGS by showing a substantial to almost perfect interobserver agreement among the six reviewers with a wide range of clinical experience (3rd-year resident to an expert in liver imaging with > 20 years of experience). Although the interobserver agreement differed according to the level of experience, over 75% of cases with discrepant grading were on the decisions between the two lowest grades, and the vast majority of the subjects with inter-observer discrepancy (> 94%) had pathologic TS < 30% and MaS < 10%. Therefore, most interobserver discrepancies seem not related to misclassifications in the assessment of significant HS from a clinical point of view, which again suggests that our VGS has an acceptable reproducibility and reliability across clinicians with a wide range of experience.

Several studies have demonstrated that weight loss achieved through diet and exercise can significantly reduce HS in potential living liver donors [24–26]. Although liver biopsy is a current reference standard for the evaluation of HS, given its drawbacks and the necessity for a repetitive assessment of liver fat content during weight loss, the feasibility of liver biopsy in potential liver donors with HS is questionable. Using the VGS with G1 as a cutoff (i.e., G0 vs. G1–3), all reviewers showed very few false-negative results to detect TS ≥ 30% and MaS ≥ 10%. This suggests the potential utility of the VGS as a screening tool to guide decision-making for liver biopsy: selective biopsy may be performed for those with a VGS of G1 or higher, in whom the possibility of clinically significant HS would not be excluded.

Our study has several limitations. First, the reference standard for HS was the liver biopsy, which may be subject to sampling bias. As a recent study has suggested high accuracy of MRI techniques such as PDFF in quantifying HS [10], comparisons between CT and MRI would result in further conclusive results, which we are planning as a subsequent study. Second, the pathologic HS in our study population was skewed towards a milder degree of HS which reflected the characteristics of the living liver donor candidates encountered in a real-world setting. It should be noted that when applying the VGS in a population with substantial discrepancies in the prevalence and profile of HS compared with our study population, the VGS could yield more false-positive or false-negative results than in our study. In such a situation, liver biopsy should remain as the reference standard. Third, the interval between CT and liver biopsy was relatively long (up to three months). However, over 90% of subjects (648/717) had an interval of < 1 month. Lastly, the VGS has limited ability to discriminate between HS and steatohepatitis, and between nonalcoholic fatty liver disease and nonalcoholic steatohepatitis, which are inherent limitations of CT scans.

In conclusion, our VGS developed for the visual assessment of HS on unenhanced CT images showed acceptable reliability and reproducibility for diagnosing clinically significant HS, especially MaS, in a large population of potential liver donors. Our VGS could be useful for the non-invasive assessment of HS in potential living liver donors during pre-donation workup.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s12072-022-10373-0.

Author contributions KWK contributed to the study's conception and design. Material preparation, data collection and analysis were performed by HJP and KWK. The first draft of the manuscript was written by HJP and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Funding This research was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Science, ICT and Future Planning (No. 2017R1E1A1A03070961).

Data availability Not applicable.

Declarations

Conflict of interest Hyo Jung Park, Kyoung Won Kim, Heon-Ju Kwon, Sunyoung Lee, Dong Wook Kim, Hye Hyeon Moon, Gi-Won Song, and Sung-Gyu Lee declare that they have no conflict of interest.

Clinical trials registration Not applicable.

Ethics approval The study protocol was approved by the medical ethics committee of Asan Medical Center.

Animal research (ethics) Not applicable.

Consent to participate (ethics) The requirement for written informed consent was waived because of the retrospective nature of this study.

Consent to publish (ethics) All authors have read and approved the manuscript and are aware that they are listed as an author on this paper.

References

1. Thyfault JP, Rector RS. Exercise combats hepatic steatosis: potential mechanisms and clinical implications. Diabetes. 2020;69(4):517–524
2. Perez-Daga JA, Santoyo J, Suárez MA, et al. Influence of degree of hepatic steatosis on graft function and postoperative complications of liver transplantation. Transplant Proc. 2006;38(8):2468–2470
3. Spitzer AL, Lao OB, Dick AA, et al. The biopsied donor liver: incorporating macrosteatosis into high-risk donor assessment. Liver Transpl. 2010;16(7):874–884
4. Miller CM, Quintini C, Dhawan A, et al. The International Liver Transplantation Society Living Donor Liver Transplant Recipient Guideline. Transplantation. 2017;101(5):938–944
5. El-Badry AM, Breitenstein S, Jochum W, et al. Assessment of hepatic steatosis by expert pathologists: the end of a gold standard. Ann Surg. 2009;250(5):691–697
6. Machado MV, Cortez-Pinto H. Non-invasive diagnosis of non-alcoholic fatty liver disease. A critical appraisal. J Hepatol. 2013;58(5):1007–1019
7. Merriman RB, Ferrell LD, Patti MG, et al. Correlation of paired liver biopsies in morbidly obese patients with suspected nonalcoholic fatty liver disease. Hepatology. 2006;44(4):874–880
8. Larson SP, Bowers SP, Palekar NA, Ward JA, Pulcini JP, Harrison SA. Histopathologic variability between the right and left lobes of the liver in morbidly obese patients undergoing Roux-en-Y bypass. Clin Gastroenterol Hepatol. 2007;5(11):1329–1332
9. Ratziu V, Charlotte F, Heurtier A, et al. Sampling variability of liver biopsy in nonalcoholic fatty liver disease. Gastroenterology. 2005;128(7):1898–1906
10. Reeder SB, Hu HH, Sirlin CB. Proton density fat-fraction: a standardized MR-based biomarker of tissue fat concentration. J Magn Reson Imaging. 2012;36(5):1011–1014
11. Zhang YN, Fowler KJ, Hamilton G, et al. Liver fat imaging-a clinical overview of ultrasound, CT, and MR imaging. Br J Radiol. 2018;91(1089):20170959
12. Xu L, Duanmu Y, Blake GM, et al. Validation of goose liver fat measurement by QCT and CSE-MRI with biochemical extraction and pathology as reference. Eur Radiol. 2018;28(5):2003–2012
13. Guo Z, Blake GM, Li K, et al. Liver fat content measurement with quantitative CT validated against MRI proton density fat fraction: a prospective study of 400 healthy volunteers. Radiology. 2020;294(1):89–97
14. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics. 1988;44(3):837–845
15. Lambert J, Halfon P, Penaranda G, Bedossa P, Cacoub P, Carrat F. How to measure the diagnostic accuracy of noninvasive liver fibrosis indices: the area under the ROC curve revisited. Clin Chem. 2008;54(8):1372–1378
16. Fluss R, Faraggi D, Reiser B. Estimation of the Youden Index and its associated cutoff point. Biom J. 2005;47(4):458–472
17. Hamer OW, Aguirre DA, Casola G, Lavine JE, Woenckhaus M, Sirlin CB. Fatty liver: imaging patterns and pitfalls. Radiographics. 2006;26(6):1637–1653
18. Ma X, Holalkere NS, Kambadakone RA, Mino-Kenudson M, Hahn PF, Sahani DV. Imaging-based quantification of hepatic fat: methods and clinical applications. Radiographics. 2009;29(5):1253–1277
19. Boas FE, Fleischmann D. CT artifacts: causes and reduction techniques. Imaging Med. 2012;4(2):229–240
20. Barrett JF, Keat N. Artifacts in CT: recognition and avoidance. Radiographics. 2004;24(6):1679–1691
21. Miyaji J, Katayama T J, Uchida A, Tanaka M, Yamazaki E. CT and MRI findings for fatty liver and differential diagnosis. In European Congress of Radiology. 2019
22. Whiting PF, Rutjes AW, Westwood ME, et al. QUADAS-2: a revised tool for the quality assessment of diagnostic accuracy studies. Ann Intern Med. 2011;155(8):529–536
23. Sica GT. Bias in research studies. Radiology. 2006;238(3):780–789
24. Doyle A, Adeyi O, Khalili K, et al. Treatment with Optifast reduces hepatic steatosis and increases candidacy rates for living donor liver transplantation. Liver Transpl. 2016;22(9):1295–1300
25. Hwang S, Lee SG, Jang SJ, et al. The effect of donor weight reduction on hepatic steatosis for living donor liver transplantation. Liver Transpl. 2004;10(6):721–725
26. Jin YJ, Kim KM, Hwang S, et al. Exercise and diet modification in non-obese non-alcoholic fatty liver disease: analysis of biopsies of living liver donors. J Gastroenterol Hepatol. 2012;27(8):1341–1347