Abdominal Radiology

© Springer Science+Business Media, LLC, part of Springer Nature 2018
Published online: 20 August 2018

Abdom Radiol (2019) 44:482–492
https://doi.org/10.1007/s00261-018-1745-3

Inter-reader agreement of magnetic resonance imaging proton density fat fraction and its longitudinal change in a clinical trial of adults with nonalcoholic steatohepatitis

Jonathan C. Hooker,[1] Gavin Hamilton,[1] Charlie C. Park,[1] Steven Liao,[1] Tanya Wolfson,[2] Soudabeh Fazeli Dehkordy,[1] Cheng William Hong,[1] Adrija Mamidipalli,[1] Anthony Gamst,[2] Rohit Loomba,[3] and Claude B. Sirlin[1]

[1]Liver Imaging Group, Department of Radiology, University of California, San Diego, 9500 Gilman Drive, San Diego, CA 92093-0888, USA

[2]Computational and Applied Statistics Laboratory (CASL), San Diego Supercomputer Center (SDSC), University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

[3]Division of Epidemiology, Department of Family Medicine and Preventive Medicine, University of California at San Diego, 9500 Gilman Drive, La Jolla, CA 92093, USA

Abstract

Purpose: To determine the inter-reader agreement of magnetic resonance imaging proton density fat fraction (PDFF) and its longitudinal change in a clinical trial of adults with nonalcoholic steatohepatitis (NASH).
Study type: We performed a secondary analysis of a placebo-controlled randomized clinical trial of a bile acid sequestrant in 45 adults with NASH. A six-echo spoiled gradient-recalled-echo magnitude-based fat quantification technique was performed at 3 T. Three independent readers measured MRI-PDFF by placing one primary and two additional regions of interest (ROIs) in each segment at both time points. Cross-sectional agreement between the three readers was evaluated using intra-class correlation coefficients (ICCs) and coefficients of variation (CV). Additionally, we used Bland–Altman analyses to examine pairwise agreement between the three readers at baseline, end of treatment (EOT), and for longitudinal change.
Results: Using all ROIs by all readers, mean PDFF at baseline, at EOT, and mean change in PDFF was 16.1%, 16.0%, and 0.07%, respectively. The 27-ROI PDFF measurements had 0.998 ICC and 1.8% CV at baseline, 0.998 ICC and 1.8% CV at EOT, and 0.997 ICC for longitudinal change. The 9-ROI PDFF measurements had corresponding values of 0.997 and 2.6%, 0.996 and 2.4%, and 0.994. Using 27 ROIs, the magnitude of the bias between readers for whole-liver PDFF measurement ranged from 0.03% to 0.06% points at baseline, 0.01% to 0.07% points at EOT, and 0.01% to 0.02% points for longitudinal change.
Conclusion: Inter-reader agreement for measuring whole-liver PDFF and its longitudinal change is high. 9-ROI measurements have only slightly lower agreement than 27-ROI measurements.

Key words: Liver—Inter-reader agreement—Quantitative imaging biomarker—Region of interest—Proton density fat fraction (PDFF)—Reproducibility

Proton density fat fraction (PDFF) is a quantitative imaging biomarker of hepatic fat content [1–5]. Confounder-corrected, chemical-shift-encoded (CSE), magnetic resonance imaging (MRI) estimates PDFF by acquiring gradient-recalled-echo images at multiple appropriately spaced echo times (TEs) to correct for R2* (1/T2*) signal decay while simultaneously measuring fat/water chemical-shift-related signal oscillation. This method also uses a low flip angle to minimize T1 bias and applies fat spectral modeling to account for the multi-peak nature of fat [6–8]. PDFF can be measured by MRI with complex reconstruction (MRI-C) [9–11] or

Correspondence to: Claude B. Sirlin; *email:* csirlin@ucsd.edu

magnitude reconstruction (MRI-M). As MRI-C is not available on all scanners, MRI-M is preferred for multicenter clinical trials. Prior studies have shown that MRI-M estimates hepatic PDFF accurately with respect to MR spectroscopy and liver biopsy [2, 9, 12–14], with high precision within the same scanner [15, 16], at different field strengths [9, 17], and across scanner platforms [17]. Due to its excellent diagnostic accuracy and reproducibility, PDFF estimated by MRI-M has been used as an endpoint in several single-center [3, 18–22] and multicenter [23, 24] clinical trials.

Whole-liver PDFF is measured by averaging the PDFF values from multiple regions of interest (ROIs) placed manually by image analysts (readers) in different parts of the liver [1, 15, 21, 23]. A multiple-ROI approach is used to avoid sampling variability because the spatial distribution of liver PDFF is nonuniform [25, 26]. Placement by readers is needed because automated methods to place ROIs are not readily available on most scanners. The optimal ROI-based sampling strategy is not yet established, with either one ROI [21, 27] or three ROIs [18, 28, 29] being placed per Couinaud segment. The exact placement was left to the readers' discretion, which is inherently subjective and may introduce a source of variability. Although inter-reader agreement for PDFF measurement has been assessed for MRI-C in adult healthy volunteers [30] and in adults with mixed liver disease [31], inter-reader agreement for PDFF measurement has not been assessed for MRI-M despite its relevance for clinical trials. Moreover, until now, reader agreement studies have focused on cross-sectional PDFF measurement in adult healthy volunteers [30] or adults with mixed liver disease [31]; the inter-reader agreement for PDFF measurement in adults with nonalcoholic steatohepatitis (NASH) and for longitudinal PDFF change is unexamined.

Therefore, the primary purpose of this study was to assess inter-reader agreement for MRI-M PDFF measurement cross-sectionally, for its longitudinal change in adults with NASH participating in a prospective clinical trial, and to examine the influence of individual readers on the results of a previously published clinical trial. A secondary purpose was to assess inter-reader agreement in individual liver segments.

Materials and methods

Study design

This was a secondary cross-sectional and longitudinal analysis of MRI-M PDFF data collected prospectively at baseline and end of treatment (EOT) as part of a clinical trial in adults with histology-confirmed NASH randomized to receive a bile acid sequestrant or placebo [18] for 24 weeks. Patients enrolled in that clinical trial were expected to undergo confounder-corrected, CSE, MRI-M at baseline and EOT, with computation of PDFF at both time points. The clinical trial and this secondary analysis were approved by an Institutional Review Board and complied with the Health Insurance Portability and Accountability Act. Patients signed informed consent to participate in the trial and to have their MRI data analyzed.

Participants

Fifty patients were enrolled in the clinical trial between January 2010 and January 2011. Inclusion criteria for the trial were 18 years of age or older, 5% or greater MRI-M PDFF at time of enrollment, and biopsy-confirmed NASH within 6 months of randomization [18]. The trial was a double-blind placebo-controlled study of a medication's effect on liver PDFF in NASH patients.

For this secondary analysis, we identified all enrolled patients that had MRI at both baseline and EOT, as this provided the opportunity to analyze inter-reader agreement cross-sectionally at two time points, as well as to analyze inter-reader agreement for longitudinal change. Thus, patients were excluded from the secondary analysis if an MRI was not performed at either time points. Demographic, anthropometric, histologic, and drug allocation data were recorded as part of the clinical trial [18]. This secondary analysis was blinded to treatment or placebo status of patients as treatment response was not relevant to its purpose.

Imaging technique

Noncontrast scans were performed at 3 T (GE Signa EXCITE HDxt, GE Healthcare, Waukesha, WI). Patients were instructed to fast for 4 h prior to being scanned head first and supine, with an eight-channel phased-array abdominal coil centered over the liver. A dielectric pad was placed between the coil and the abdominal wall to reduce shading from B1 heterogeneity. A six-echo spoiled gradient-recalled-echo magnitude-based fat quantification technique was performed with parameters listed in Table 1. A low flip angle was used to minimize T1 bias [32]. Echoes were acquired at six nominally out-of-phase and in-phase TEs from 1.15 to 6.9 ms to permit correction for R2* signal decay and chemical-shift-based separation of fat and water signals, assuming water to be the dominant component in the liver. Field of view and matrix size were adjusted to ensure the entire liver was imaged within one breath-hold, although there were two cases where two breath-holds were required to image the entire liver. Parallel imaging was not used as the version available at the time was prone to artifact, especially in obese adults.

Postprocessing

Parametric PDFF maps were generated by the scanner computer using a custom algorithm [6, 12] applied to the

Table 1. Acquisition parameters for MRI-PDFF estimation technique

Parameters	Values
MR acquisition type	2D
Repetition time (ms)	150–225
Number of echoes	6
Echo time for each echo (ms)	1.15, 2.30, 3.45, 4.60, 5.75, 6.90
Slice thickness (mm)	8–11
Inter-slice gap (mm)	0
Number of slices	17–34 (median 23)
Number of averages	1
Acquisition matrix	224 × 160–160 × 128
Flip angle	10°
Field of view (mm)	360–440
Phase field of view (%)	65–90
Acquisition time/breath-hold (s)	16–34
Number of breath-holds for whole-liver coverage	1–2

source images pixel by pixel. This algorithm assumes exponential R2* signal decay across the six TEs. It models water as a single peak at 4.7 ppm with fat being a multi-component signal [6, 7] with relative amplitudes of 0.047, 0.039, 0.006, 0.120, 0.700, and 0.088 at 5.30, 4.20, 2.75, 2.10, 1.30, and 0.90 ppm, respectively. Source images and parametric maps were transferred for offline analysis.

Readers

For this study, three independent readers not involved in the initial clinical trial analyzed the images. One reader (J.C.H.) had 1 year of experience in MRI-M PDFF analysis ("experienced" reader), while the other two readers (C.C.P. and S.L.) were medical students who received 8 h of PDFF analysis training by the senior reader and 30 h of supervised practice ("novice" readers). The training and practice sessions were done on imaging datasets from patients not involved in this clinical trial.

Image analysis

Each reader was instructed to place three circular ROIs (1 cm radius) in each of the nine segments of the liver (a total of 27 ROIs) on each baseline and each EOT scan (in a paired fashion), while avoiding edges of the liver, segmental boundaries, vessels, and artifacts. 27 ROIs were used to match the analysis done in the clinical trial [18]. ROIs of this size and shape were selected to match the methods reported in recent clinical trials using MRI-PDFF as an endpoint [19, 21, 23, 24]. The first ROI in each segment was placed as centrally as possible within the segment. The readers understood that this central ROI would be considered primary, to be taken as the single ROI in the segment if they were doing a nine-ROI per liver analysis. The remaining two ROIs per segment were considered secondary; these were placed in a similar location as the primary ROI on the slices immediately above and below if the segment was large enough to allow this, and on the same slice otherwise.

ROIs were placed on source images rather than PDFF maps to reduce feedback bias. The readers were also instructed to judge individually, among the six source images at different TEs, the best source image for ROI placement based on the image that most clearly demonstrated the hepatic anatomy. In most cases, this was the fifth echo (TE = 5.75 ms) image, although in some cases an earlier out-of-phase echo image was used if it was determined by the reader to provide better anatomic visualization. The echo chosen was not recorded. Each reader took between 1 and 3 weeks to complete ROI placement for all exams.

After placing all 27 ROIs on the baseline scans, ROIs were placed on the same locations of the corresponding source images on the EOT scans using anatomic landmarks. After verifying visually that the ROIs were co-localized at the two points, the ROIs were propagated to the corresponding PDFF maps, and the PDFF values for each ROI were exported to a database. ROIs were assigned identifiers by segment and whether they were considered primary or secondary. Figure 1 illustrates the ROI placement on the baseline PDFF maps of one patient by all three readers.

The readers were blinded to each other's results, to the results of the original clinical trial, and to whether patients were assigned to drug or placebo, but were not blinded to whether examinations were baseline or EOT.

Identifying ROIs with outlier values and reviewing ROI placement

The experienced reader identified outlier PDFF values in the dataset using the following approach: for each exam, the mean of all 27 ROIs from all three readers was averaged (81 ROIs total), and the difference between this mean, and each individual ROI PDFF value was taken. An absolute difference of more the 5% was then used to define outliers. A difference of 5% was chosen based on the segmental PDFF variation reported in previous studies [25, 26, 33]. The experienced reader then visually examined all ROIs with outlier PDFFs for incorrect placement (outside the liver, outside the intended segment, on a vessel large enough or artifact severe enough that in experienced reader's judgment should be excluded from the ROI). Placement errors were recorded, but ROIs were not moved even if placed incorrectly (i.e., the original values were used in the analyses).

Statistical analyses

Study patients were summarized descriptively. Cross-sectional agreement between the three readers was evaluated using intra-class correlation coefficients (ICCs)

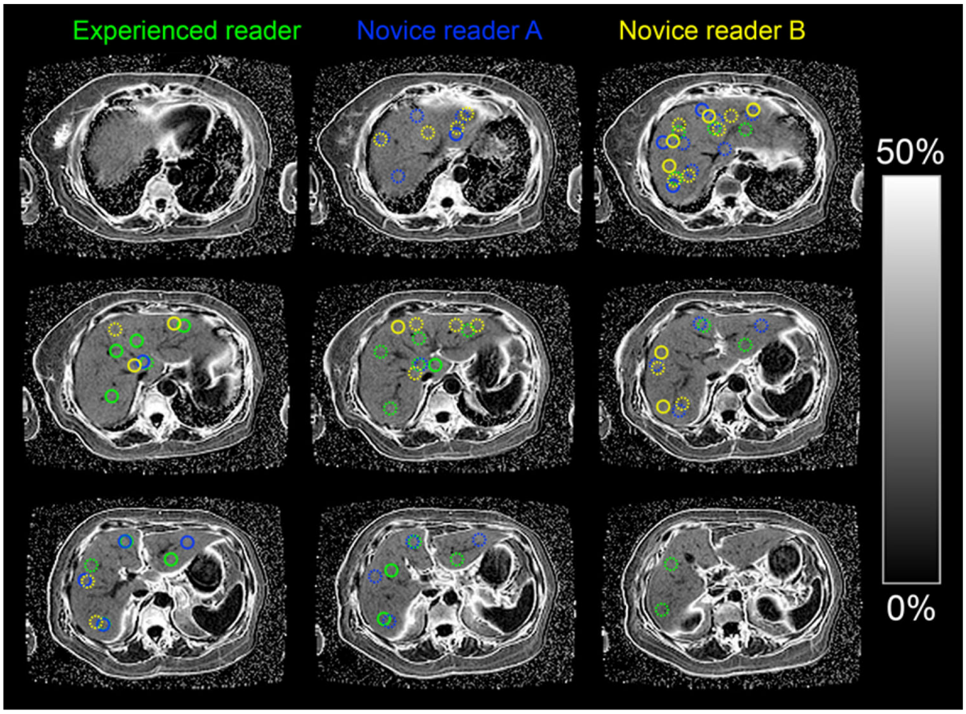

Fig. 1. The liver covered from dome (top left) to tip (bottom right) on the magnitude-based MRI proton density fat fraction (PDFF) maps. The 27 regions of interest (ROIs) placed by each reader are shown, color coded by reader. The primary ROI in each segment has a continuous border. The secondary ROIs have dashed borders. Note that the scale bar reflects the fat–water signal dominance ambiguity using magnitude reconstruction.

and coefficients of variation (CV) for dependent data, both with bootstrap-based 95% confidence intervals. These agreement metrics were computed at baseline and EOT for each segment based on the primary ROI, for the average of nine segments based on the primary ROI, and for the average of all 27 ROIs. Longitudinal agreement between the three readers was evaluated using only ICC with 95% confidence intervals; CV was not computed as it is problematic for assessing measures roughly symmetric around 0, such as longitudinal PDFF changes in a clinical trial which may be positive or negative.

Additionally, we used Bland–Altman analyses to examine pairwise agreement between the three readers for 9-ROI and 27-ROI whole-liver PDFF measurement at baseline, EOT, and for longitudinal change. Bland–Altman plots were generated. Bland–Altman biases and their *p* values, standard deviations, and 95% limits of agreement (LOA) were computed. *p* values were not adjusted for multiple comparisons to maximize sensitivity for detecting potential biases between readers. Additionally, CVs for dependent data were computed for each pairwise comparison at baseline and at EOT. CVs for PDFF change were not calculated for the reason explained above.

To explore the impact on reader agreement of misplaced ROIs (outside the liver or on a large vessel), we first identified the segment with the most outlier ROI values due to misplacement and then recomputed the ICCs and CVs for that segment after removing those misplaced ROIs. We selected the segment with the most misplaced ROIs to provide an upper bound estimate; the impact of outliers would be smaller for any other segment and for whole-liver (9-ROI and 27-ROI) PDFF measurements.

Finally, we have repeated one of the main analyses of the clinical trial: a comparison of pre- to post-treatment changes in PDFF between the treatment and placebo groups, using the PDFF results from each of the three readers.

Results

Patient characteristics

Fifty patients were enrolled in the parent clinical trial, of whom 45 (21 male, 24 female) completed both baseline and EOT MRI exams and were included in our analysis. Mean (± standard deviation) age was 47.9 ± 11.3 years. Mean body mass index was 31.3 ± 4.9 kg/m^2. Using all ROIs from all readers, mean whole-liver PDFF was 16.1% (range 4.7%–33.8%) at baseline and 16.0% (range 4.5%–30.5%) at EOT. Mean change in whole-liver PDFF was − 0.07% ± 6.41% (range − 17.35% to 12.06%). Using only the primary ROIs from all readers, mean individual-segment PDFF values ranged from 14.8% (segment 2) to 16.8% (segment 8) at baseline and from 14.7% (segment 2) to 16.6% (segments 7, 8) at EOT.

Table 2. Average PDFF values (expressed as percentages) at baseline and end of treatment across all ROIs for each reader and overall

	Average PDFF (%) at baseline for each reader				Average PDFF (%) at end of treatment for each reader			
	Exp. Reader	Novice Reader A	Novice Reader B	Overall	Exp. Reader	Novice Reader A	Novice Reader B	Overall
Segment 1	15.5	15.4	16.1	15.7	15.5	15.4	15.9	15.6
Segment 2	14.5	14.8	15.2	14.8	14.2	15.0	14.8	14.7
Segment 3	15.8	15.5	15.7	15.7	15.9	15.4	15.7	15.7
Segment 4a	16.3	16.5	16.6	16.5	16.1	16.4	16.3	16.3
Segment 4b	16.4	16.3	16.7	16.5	16.2	16.2	16.4	16.3
Segment 5	16.3	15.8	16.3	16.1	16.0	15.8	15.9	15.9
Segment 6	15.6	15.9	15.9	15.8	15.9	16.1	16.1	16.0
Segment 7	16.9	16.6	16.4	16.6	16.8	16.6	16.5	16.6
Segment 8	17.1	16.9	16.5	16.8	16.8	16.5	16.3	16.6
9 ROI average	16.1	16.0	16.1	16.1	16.0	15.9	16.0	16.0
27 ROI average	16.0	16.1	16.1	16.1	16.0	16.0	16.0	16.0

Exp. Reader, experienced reader

Table 3. Average PDFF change values (expressed as absolute differences in percentage points) from baseline to end of treatment across all ROIs for each reader and overall

	Average change in PDFF (%) for each reader			
	Exp. Reader	Novice Reader A	Novice Reader B	Overall
Segment 1	− 0.01	− 0.06	− 0.16	− 0.08
Segment 2	− 0.46	0.21	− 0.34	− 0.19
Segment 3	0.04	− 0.13	0.01	− 0.03
Segment 4a	− 0.32	− 0.12	− 0.3	− 0.25
Segment 4b	− 0.22	− 0.1	− 0.25	− 0.19
Segment 5	− 0.28	0.05	− 0.42	− 0.22
Segment 6	0.28	0.26	0.17	0.24
Segment 7	− 0.38	0.02	0.17	− 0.06
Segment 8	− 0.38	− 0.34	− 0.17	− 0.3
9 ROI average	− 0.1	− 0.02	− 0.14	− 0.09
27 ROI average	− 0.07	− 0.09	− 0.06	− 0.07

Exp. Reader, experienced reader

Mean change in individual-segment PDFF varied from − 0.2% (segments 4a, 4b, 5, 8) to + 0.2% (segment 6). Tables 2 and 3 summarize the whole-liver and individual-segment PDFF values at baseline and EOT as well as their longitudinal changes.

Inter-reader agreement for measuring PDFF cross-sectionally

ICCs and CVs for whole-liver (9- and 27-ROI averages) and individual-segment PDFF measurements at baseline and at EOT are summarized in Table 4.

Whole liver

At baseline, the ICC for the 9-ROI whole-liver PDFF measurement was 0.997 while the CV was 2.6%. The ICC for the 27-ROI whole-liver PDFF measurement was 0.998 while the CV was 1.8%. At EOT, the ICC for the 9-ROI whole-liver PDFF measurement was 0.996 while the CV was 2.4%. The ICC for the 27-ROI whole-liver PDFF measurement was 0.998 while the CV was 1.8%.

Individual segment

At baseline, the ICCs for primary-ROI single-segment PDFF measurements ranged from 0.957 to 0.990 while CVs ranged from 9.5% to 4.9%, depending on the segment. At EOT, the ICCs for primary-ROI single-segment PDFF measurements ranged from 0.943 to 0.992 while CVs ranged from 10.8% to 3.7%, depending on the segment. Segments 7 and 8 nominally had the highest ICCs at baseline (0.988 and 0.990, respectively) and EOT (0.992 and 0.988, respectively) as well as the lowest CVs at baseline (5.1% and 4.9%, respectively) and EOT (3.7% and 4.7%, respectively). In contrast, segments 1 and 2 nominally had the lowest ICCs at baseline (0.965 and 0.957, respectively) and EOT (0.955 and 0.943) as well as the highest CVs at baseline (8.6% and 9.5%, respectively) and EOT (9.0% and 10.8%, respectively).

Inter-reader agreement for measuring PDFF change longitudinally

ICCs for whole-liver (9- and 27-ROI averages) and individual-segment PDFF-change measurements are summarized in Table 4.

Table 4. Agreement between readers at baseline, at end of treatment (EOT), and for longitudinal change at whole-liver level using 27 ROIs, at whole-liver level using 9 ROIs, and at segmental level using 1 ROI (the primary ROI)

Segments	Baseline ICC [95% CI]	EOT ICC [95% CI]	Baseline CV (%) [95% CI]	EOT CV (%) [95% CI]	Change ICC [95% CI]
1	0.965 [0.928–0.982]	0.955 [0.866–0.980]	8.6 [6.6–11.5]	9.0 [6–14.8]	0.951 [0.898–0.977]
2	0.957 [0.911–0.978]	0.943 [0.886–0.970]	9.5 [7.2–13]	10.8 [8.3–16.6]	0.893 [0.760–0.950]
3	0.974 [0.949–0.986]	0.964 [0.939–0.980]	7.6 [6.1–10]	8.2 [6.7–10.1]	0.954 [0.924–0.972]
4a	0.986 [0.974–0.992]	0.977 [0.964–0.985]	5.1 [4.1–6.5]	6.3 [5.2–8]	0.969 [0.937–0.984]
4b	0.983 [0.968–0.990]	0.980 [0.966–0.988]	5.9 [4.7–7.6]	6.0 [4.8–7.8]	0.970 [0.948–0.983]
5	0.979 [0.969–0.987]	0.982 [0.965–0.990]	6.8 [5.5–8.1]	6.1 [4.8–7.9]	0.977 [0.959–0.987]
6	0.988 [0.975–0.993]	0.985 [0.973–0.992]	5.3 [4.3–6.8]	5.3 [4.2–7]	0.982 [0.970–0.990]
7	0.988 [0.979–0.993]	0.992 [0.987–0.995]	5.1 [4.2–6.5]	3.7 [3.1–4.6]	0.988 [0.980–0.993]
8	0.990 [0.981–0.995]	0.988 [0.980–0.993]	4.9 [3.8–6.4]	4.7 [3.9–6]	0.977 [0.959–0.986]
9 ROIs	0.997 [0.994–0.998]	0.996 [0.995–0.998]	2.6 [2.1–3.4]	2.4 [2–2.9]	0.994 [0.991–0.997]
27 ROIs	0.998 [0.997–0.999]	0.998 [0.997–0.999]	1.8 [1.4–2.4]	1.8 [1.5–2.2]	0.997 [0.995–0.998]

Bootstrap-based 95% confidence intervals are in brackets
ICC, intra-class correlation; CV, coefficient of variation for dependent data

Table 5. Summary of Bland–Altman analyses

Whole-liver PDFF	Reader pair	Bias (%)	SD (%)	*p* values	Limits of agreement		CV (%)
					Lower bound (%)	Upper bound (%)	
Baseline							
9 ROIs	Experienced vs. Novice-A	0.09	0.59	0.32	− 1.07	1.24	2.60
	Experienced vs. Novice-B	− 0.08	0.67	0.45	− 1.39	1.24	2.93
	Novice-A vs. Novice-B	− 0.16	0.44	0.02	− 1.02	0.70	2.04
27 ROIs	Experienced vs. Novice-A	− 0.03	0.40	0.63	− 0.82	0.76	1.76
	Experienced vs. Novice-B	− 0.06	0.43	0.36	− 0.91	0.79	1.90
	Novice-A vs. Novice-B	− 0.03	0.39	0.60	− 0.79	0.73	1.68
End of treatment							
9 ROIs	Experienced vs. Novice-A	0.01	0.55	0.89	− 1.06	1.08	2.39
	Experienced vs. Novice-B	− 0.03	0.59	0.71	− 1.18	1.12	2.57
	Novice-A vs. Novice-B	− 0.05	0.54	0.58	− 1.10	1.01	2.37
27 ROIs	Experienced vs. Novice-A	− 0.01	0.44	0.85	− 0.88	0.85	1.93
	Experienced vs. Novice-B	− 0.07	0.42	0.29	− 0.88	0.75	1.84
	Novice-A vs. Novice-B	− 0.05	0.39	0.36	− 0.82	0.72	1.73
Change from baseline to end of treatment							
9 ROIs	Experienced vs. Novice-A	− 0.08	0.67	0.45	− 1.38	1.23	NA
	Experienced vs. Novice-B	0.04	0.79	0.71	− 1.48	1.57	NA
	Novice-A vs. Novice-B	0.12	0.64	0.22	− 1.13	1.37	NA
27 ROIs	Experienced vs. Novice-A	0.02	0.52	0.83	− 1.01	1.04	NA
	Experienced vs. Novice-B	− 0.01	0.50	0.93	− 0.99	0.98	NA
	Novice-A vs. Novice-B	− 0.02	0.46	0.73	− 0.92	0.87	NA

SD, standard deviation of the bias; CV, coefficient of variation; NA, not applicable

The ICC for the 9-ROI whole-liver PDFF-change measurement was 0.994. The ICC for the 27-ROI whole-liver PDFF-change measurement was 0.997. The ICC for primary-ROI single-segment PDFF-change measurements ranged from 0.893 to 0.988, depending on the segment. Segments 6 and 7 nominally had the highest ICCs (0.982 and 0.988, respectively), while segments 1 and 2 nominally had the lowest (0.951 and 0.893, respectively).

Bland–Altman analyses

The results of the Bland–Altman analysis for pairwise inter-reader agreement are summarized in Table 5.

Using 9 ROIs, the magnitude of the bias between readers for whole-liver PDFF measurement ranged from 0.08% to 0.16% points at baseline, 0.01% to 0.05% points at EOT, and 0.04% to 0.12% points for longitudinal change. CVs ranged from 2.04% to 2.93% at baseline and 2.37% to 2.57% at EOT.

Using 27 ROIs, the magnitude of the bias between readers for whole-liver PDFF measurement ranged from 0.03% to 0.06% points at baseline, 0.01% to 0.07% points at EOT, and 0.01% to 0.02% points for longitudinal change. CVs ranged from 1.68% to 1.90% at baseline and 1.73% to 1.93% at EOT.

Of the 18 pairwise inter-reader bias comparisons, only one was significant: the 0.16%-point bias between the two novice readers for baseline PDFF using 9 ROIs.

Figure 2 shows Bland–Altman plots for baseline and for longitudinal change measurements. EOT data plots (not shown) are similar to the baseline data plots.

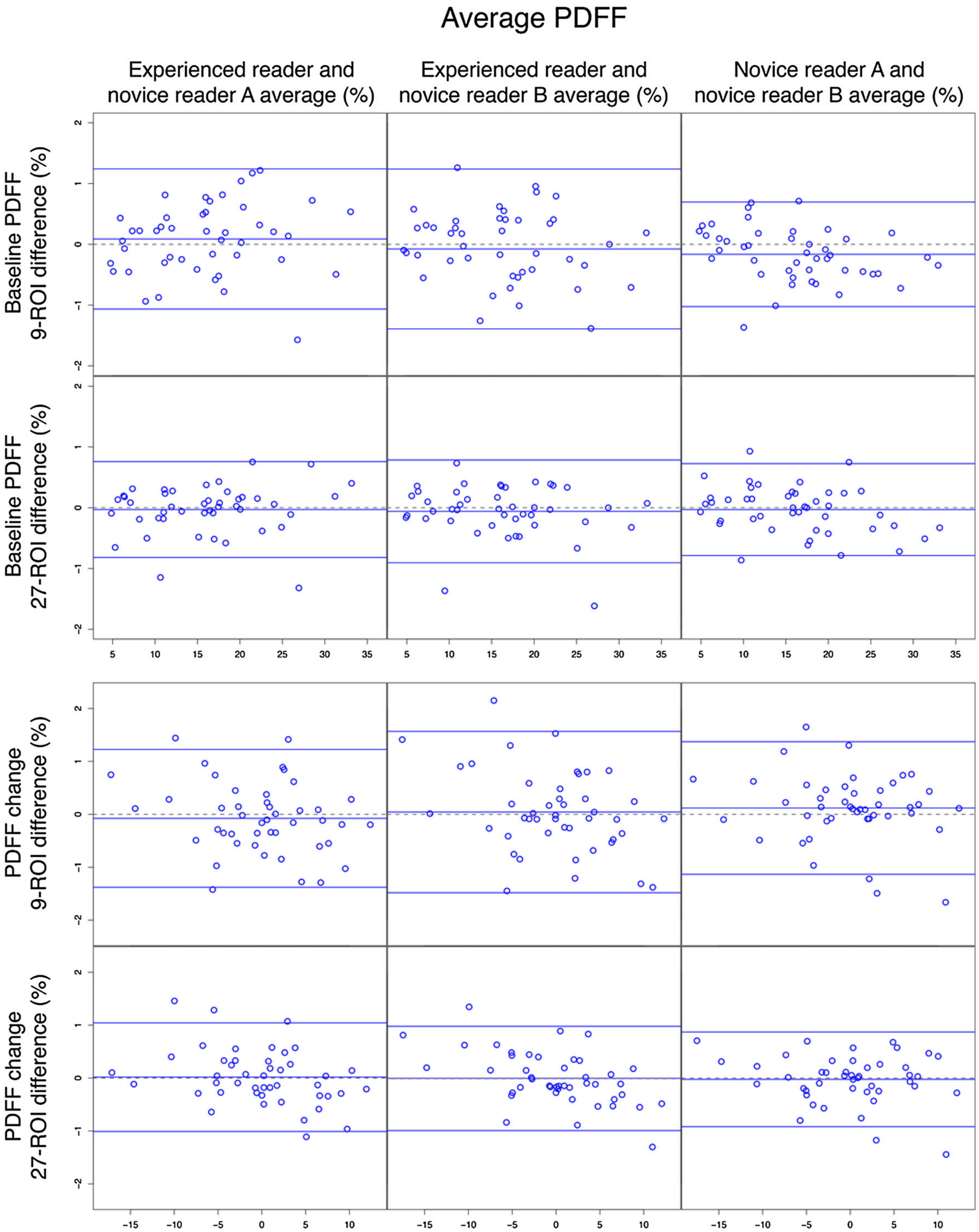

Fig. 2. The results of the Bland–Altman analysis for measuring PDFF at baseline and its change longitudinally for 9- and 27-ROI methods. Plots for measuring PDFF at end of treatment (not shown) are similar to those at baseline.

ROIs with outlier PDFF values

Out of 7290 ROIs placed (45 patients × 2 time points × 27 ROIs per patient × 3 readers), 241 (3.3%) ROIs had outlier PDFF values. Most outlier ROIs (186/241, 77%) were verified as being appropriately positioned on retrospective review by the experienced reader, being entirely surrounded by liver tissue, being entirely contained within the intended segment, and not overlying a major blood vessel or artifact.

The other 55/241 (23%) were improperly positioned (i.e., 0.7% of all ROIs, 55/7290). For each reader, the most common type of misplacement was inclusion of

Table 6. Summary of outliers and source of error for each reader and the totals

	Reader			Total
	Experienced	Novice A	Novice B	
Number of outliers	77	92	72	241
Incorrectly placed ROIs	11	33	11	55
Cause				
Outside liver	9	25	7	41
Outside segment	0	2	2	4
On a blood vessel	2	6	1	9
On an artifact	0	0	1	1
Segments				
1	0	5	2	7
2	6	10	5	21
3	1	5	1	7
4a	1	4	0	5
4b	2	3	0	5
5	0	3	1	4
6	0	3	2	5
7	1	0	0	1
8	0	0	0	0

tissue outside the liver (Table 6); 38 of the 55 (69%) incorrectly placed ROIs were in the left lobe of the liver, 10 (18%) in the right lobe, and the remaining 7 (13%) in the caudate lobe.

Segment 2 had the most outliers ($N = 21$): 14 of the 21 (67%) were placed in the heart and 7 of the 21 (33%) were placed on a major blood vessel; 3 of the misplaced ROIs were considered primary by the reader. In a post hoc analysis of the primary ROI for segment 2, ICCs improved from 0.957 to 0.970 at baseline, 0.943 to 0.947 at EOT, and 0.893 to 0.943 for longitudinal change; CVs improved from 9.5% to 8.1% and 10.8% to 10.5% at baseline and EOT, respectively.

Inter-reader agreement on results of clinical trial

The result of the clinical trial was confirmed by each reader, with mean differences in PDFF change between treatment and placebo groups of 5.4%–5.8%, similar to the 5.6% mean difference reported by the clinical trial [18] The differences between change in treatment and change in placebo groups were significant at all segments and ROI averages for each reader. Table 7 details the differences for each reader in each segment, using an average of 9 ROIs, and using an average of 27 ROIs.

Discussion

Key findings

In this secondary analysis of adults with NASH participating in a clinical trial, we found that inter-reader agreement for whole-liver and segment-level PDFF measurement using MRI-M was high cross-sectionally at both baseline and EOT as well as longitudinally. Using three readers of varying experience, both the 9- and the 27-ROI methods achieved inter-reader ICCs > 0.995 and CVs $< 3\%$ for measuring PDFF cross-sectionally as well as ICCs > 0.990 for measuring PDFF change longitudinally. In pairwise comparisons of readers, all Bland–Altman biases were $< 0.2\%$ points in magnitude and all CVs were $< 3\%$. Compared to the 9-ROI method, the 27-ROI method provided slightly higher ICC point estimates and slightly narrower CV point estimates. With one exception, all segment-level PDFF cross-sectional measurements achieved ICCs > 0.950 and CVs $< 10\%$ for measuring PDFF cross-sectionally at both time points as well as ICCs > 0.950 for measuring PDFF change longitudinally. The exception, segment 2, had the most ROIs with outlier PDFF values and the most misplaced ROIs, which may explain its wider reader variability. Removing the misplaced ROIs in a post hoc analysis improved the ICCs and CVs for segment 2, but it continued to have wider reader variability than almost every other segment. Segments 1 and 2 also had relatively low inter-reader reproducibility. Thus, the segments with the widest reader variability were in the caudate or in the left lateral sector.

Table 7. Results of clinical trial as determined by each reader

	Exp. Reader average difference	Exp. Reader *p* values	Novice Reader A average difference	Novice Reader A *p* values	Novice Reader B average difference	Novice Reader B *p* values
Segment 1	5.0	0.0071	5.5	0.0042	5.5	0.0033
Segment 2	4.1	0.0182	4.7	0.0119	6.3	0.0018
Segment 3	5.5	0.0042	5.8	0.0044	6.6	0.0012
Segment 4a	4.8	0.0068	5.9	0.0022	5.8	0.0023
Segment 4b	5.0	0.0067	5.6	0.0047	5.5	0.0057
Segment 5	5.0	0.0102	5.9	0.0033	5.3	0.0090
Segment 6	5.8	0.0026	5.8	0.0050	5.8	0.0034
Segment 7	6.1	0.0010	6.2	0.0007	6.2	0.0015
Segment 8	5.2	0.0062	5.5	0.0031	5.6	0.0031
9 ROI average	5.3	0.0037	5.6	0.0026	5.9	0.0021
27 ROI average	5.4	0.0031	5.6	0.0031	5.8	0.0024

Reproducing results of the clinical trial

For each of the readers, we found the difference between PDFF changes in treatment and placebo groups to be similar to those found in the clinical trial [18], between a 4% and 6% difference depending on the segments. The pattern of significances was similar across segments and the main (overall) result of the trial was confirmed by each reader. Minor discrepancies between individual readers had no effect on the conclusion of the clinical trial.

In context of the existing literature

Previous studies have demonstrated the accuracy [2, 8, 32, 33], repeatability [15, 16, 26, 33], and reproducibility of MRI-PDFF across different field strengths (1.5 and 3.0 T) [9, 17], scanner platforms [17, 30], and body habitus [5, 9], but until recently, the reproducibility across readers was unexamined. Two new studies assessed inter-reader agreement cross-sectionally with MRI-C PDFF. Using data from 24 adult volunteers (mean PDFF 4.87%, standard deviation 4.59% for the GE scanner) scanned on multiple scanner platforms, Serai et al. [30] reported inter-reader agreement ICC values from 0.966 to 0.995 among five readers for measuring MRI-C PDFF. Campo et al. [31] evaluated the inter-reader agreement of different ROI sizes and numbers (up to 9) for MRI-C PDFF estimation in a composite cohort (mean PDFF 5.9%, SD 8.8%) comprising 19 healthy subjects, 34 adults with a clinical indication for abdominal MRI, and 37 adults with suspected iron overload. They found that inter-reader agreement improved as the liver sampling area increased through the use of larger and/or more ROIs. With 9 ROIs comparable in size to ours, the Bland–Altman LOA were 1.5%. Our results extend the findings of these two studies to a clinical population of adults with histology-confirmed NASH with a wider range of PDFF values assessed by a magnitude reconstruction technique comparable to that used in several recent clinical trials [18, 19, 21, 27–29]. Moreover, our study examines the inter-reader agreement not only for cross-sectional measurements but also for measuring longitudinal PDFF change, which is relevant for applying MRI-PDFF as an endpoint in clinical trials.

Although our study did not compare MRI to other methods, the inter-reader agreement reported here for MRI-M compares favorably to the inter-reader agreement among pathologists for scoring steatosis on biopsy (κs ranging from 0.62 to 0.88 [34–36], or an ICC of 0.65 [37]) and for radiologists in scoring steatosis on ultrasound (κ of 0.54) [38]. This may be because of the qualitative and subjective nature of these other scoring systems.

Implications

The high agreement between readers for measuring PDFF cross-sectionally and its change longitudinally suggests that multiple readers can be used to measure PDFF in clinical trials and for clinical care. A 9-ROI sampling approach generally suffices as reader agreement remains relatively high, with much lower reader burden than the 27-ROI approach. We do not recommend a single-ROI approach, but if such an approach is applied, we recommend avoiding segments 1–3, which had the worst reader agreement. Our study was not designed to determine the causes of reader variability. Plausible explanations for the lower reproducibility in segments 1–3 are that these segments have higher spatial variability in PDFF distribution as reported by Bonekamp et al. [25]; are smaller, which makes it more difficult to place ROIs within liver boundaries while avoiding vessels and artifacts; and are more prone to having artifacts from heart motion. Additionally, source images are acquired with low flip angle to reduce T1 weighting. While the resulting images are suitable for fat quantification, they provide little contrast between organs. As a result, portions of the liver (especially segment 2) can blend with adjacent structures such as the heart and spleen. This can cause malpositioning of ROIs outside the liver. To avoid these errors, readers should be trained on recognizing the liver boundaries, especially around the caudate and left lateral sector, even when the interfaces between the liver and other organs are not clearly visible. We speculate that using multiplanar reformats may help ROI placement, but this was not tested in this study.

Limitations

Our study has several limitations. This was a single-center study performed on a single scanner at a tertiary academic center with expertise in the performance of the applied CSE–MRI technique, which may limit study generalizability to cohorts in other geographical regions or to other centers. Another limitation is that we did not assess the effect of ROI size on inter-reader agreement, although the ROI size used in this study is commonly used in clinical settings and has been used in at least nine prior published studies on PDFF [1, 2, 9, 13, 15, 16, 39–41]. Additionally, two of the readers were trained by the experienced reader, which may have introduced training bias and underestimation of inter-reader variability. Finally, intra-reader agreement was not assessed. Although intra-reader agreement is likely to be similar or higher, a direct comparison of intra- and inter-reader agreement would help determine the proportion of reader variability that is attributable to different readers. Hong and colleagues recently suggested that a 4-ROI selection strategy with two ROIs in the left lobe and two in the right lobe is a reasonable compromise between reader burden and comprehensive liver sampling; future work is needed to assess the inter-reader agreement for the 4-ROI approach. We also did not assess agreement for longitudinal change when one reader measures PDFF at baseline and an another at EOT. Finally, we did not re-review all 7290 ROIs for misplacement, only the 241 with outlier values as defined.

Conclusion

In conclusion, our study demonstrates high inter-reader agreement for measuring hepatic PDFF cross-sectionally and its change longitudinally. This suggests that when using the 9-ROI and 27-ROI methods in clinical and research settings, inter-reader agreement is sufficiently high to allow for multiple readers without meaningful variations in PDFF measurement.

Compliance with ethical standards

Funding The project described was partially supported by the National Institutes of Health Grants K23-DK090303, R01 DK083380-01, and T32EB005970. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Conflict of interest Jonathan C. Hooker, Dr. Gavin Hamilton, Charlie C. Park, Dr. Steven Liao, Tanya Wolfson, Dr. Soudabeh Fazeli Dehkordy, Dr. Cheng William Hong, Dr. Adrija Mamidipalli, and Dr. Anthony Gamst declare that they have no conflict of interest. Dr. Rohit Loomba has the following disclosure related to this study: Daiichi-Sankyo. Dr. Claude B. Sirlin has the following disclosures not related to this study. Industry service agreements and consultation agreements: Alexion, AstraZeneca, Bioclinica, BMS, Fibrogen, Galmed, Genzyme, Gilead, Fibrogen, Icon, Intercept, Isis, Janssen, NuSirt, Perspectum, Pfizer, Profil, Sanofi, Shire, Synageva, Tobira, Takeda, VirtualScopics.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the Institutional and/or National Research Committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Tang A, Tan J, Sun M, et al. (2013) Nonalcoholic fatty liver disease: MR imaging of liver proton density fat fraction to assess hepatic steatosis. Radiology 267(2):422–431
2. Tang A, Desai A, Hamilton G, et al. (2015) Accuracy of MR imaging-estimated proton density fat fraction for classification of dichotomized histologic steatosis grades in nonalcoholic fatty liver disease. Radiology 274(2):416–425
3. Schwimmer JB, Middleton MS, Behling C, et al. (2015) Magnetic resonance imaging and liver histology as biomarkers of hepatic steatosis in children with nonalcoholic fatty liver disease. Hepatology 61(6):1887–1895
4. Idilman IS, Aniktar H, Idilman R, et al. (2013) Hepatic steatosis: quantification by proton density fat fraction with MR imaging versus liver biopsy. Radiology 267(3):767–775
5. Rehm JL, Wolfgram PM, Hernando D, et al. (2015) Proton density fat-fraction is an accurate biomarker of hepatic steatosis in adolescent girls and young women. Eur Radiol 25(10):2921–2930
6. Bydder M, Yokoo T, Hamilton G, et al. (2008) Relaxation effects in the quantification of fat using gradient echo imaging. Magn Reson Imaging 26(3):347–359
7. Hamilton G, Yokoo T, Bydder M, et al. (2011) In vivo characterization of the liver fat $^{(1)}$H MR spectrum. NMR Biomed 24(7):784–790
8. Reeder SB, Cruite I, Hamilton G, Sirlin CB (2011) Quantitative assessment of liver fat with magnetic resonance imaging and spectroscopy. J Magn Reson Imaging 34(4):spcone
9. Artz NS, Haufe WM, Hooker CA, et al. (2015) Reproducibility of MR-based liver fat quantification across field strength: same-day comparison between 1.5 T and 3 T in obese subjects. J Magn Reson Imaging 42(3):811–817
10. Reeder SB, McKenzie CA, Pineda AR, et al. (2007) Water–fat separation with IDEAL gradient-echo imaging. J Magn Reson Imaging 25(3):644–652
11. Yu H, Shimakawa A, Hines CD, et al. (2011) Combination of complex-based and magnitude-based multiecho water–fat separation for accurate quantification of fat-fraction. Magn Reson Med 66(1):199–206
12. Yokoo T, Bydder M, Hamilton G, et al. (2009) Nonalcoholic fatty liver disease: diagnostic and fat-grading accuracy of low-flip-angle multiecho gradient-recalled-echo MR imaging at 1.5 T. Radiology 251(1):67–76
13. Yokoo T, Shiehmorteza M, Hamilton G, et al. (2011) Estimation of hepatic proton-density fat fraction by using MR imaging at 3.0 T. Radiology 258(3):749–759
14. Permutt Z, Le TA, Peterson MR, et al. (2012) Correlation between liver histology and novel magnetic resonance imaging in adult patients with non-alcoholic fatty liver disease—MRI accurately quantifies hepatic steatosis in NAFLD. Aliment Pharmacol Ther 36(1):22–29
15. Negrete LM, Middleton MS, Clark L, et al. (2014) Inter-examination precision of magnitude-based MRI for estimation of segmental hepatic proton density fat fraction in obese subjects. J Magn Reson Imaging 39(5):1265–1271
16. Tyagi A, Yeganeh O, Levin Y, et al. (2015) Intra- and inter-examination repeatability of magnetic resonance spectroscopy, magnitude-based MRI, and complex-based MRI for estimation of hepatic proton density fat fraction in overweight and obese children and adults. Abdom Imaging 40(8):3070–3077
17. Kang GH, Cruite I, Shiehmorteza M, et al. (2011) Reproducibility of MRI-determined proton density fat fraction across two different MR scanner platforms. J Magn Reson Imaging 34(4):928–934
18. Le TA, Chen J, Changchien C, et al. (2012) Effect of colesevelam on liver fat quantified by magnetic resonance in nonalcoholic steatohepatitis: a randomized controlled trial. Hepatology 56(3):922–932
19. Cui J, Philo L, Nguyen P, et al. (2016) Sitagliptin vs. placebo for non-alcoholic fatty liver disease: a randomized controlled trial. J Hepatol 65(2):369–376
20. Tang A, Rabasa-Lhoret R, Castel H, et al. (2015) Effects of insulin glargine and liraglutide therapy on liver fat as measured by magnetic resonance in patients with type 2 diabetes: a randomized trial. Diabetes Care 38(7):1339–1346
21. Loomba R, Sirlin CB, Ang B, et al. (2015) Ezetimibe for the treatment of nonalcoholic steatohepatitis: assessment by novel magnetic resonance imaging and magnetic resonance elastography in a randomized trial (MOZART trial). Hepatology 61(4):1239–1250
22. Meisamy S, Hines CD, Hamilton G, et al. (2011) Quantification of hepatic steatosis with T1-independent, T2-corrected MR imaging with spectral modeling of fat: blinded comparison with MR spectroscopy. Radiology 258(3):767–775
23. Middleton MS, Heba ER, Hooker CA, et al. (2017) Agreement between magnetic resonance imaging proton density fat fraction measurements and pathologist-assigned steatosis grades of liver biopsies from adults with nonalcoholic steatohepatitis. Gastroenterology 153(3):753–761
24. Middleton MS, Van Natta ML, Heba ER, et al. (2017) Diagnostic accuracy of magnetic resonance imaging hepatic proton density fat fraction in pediatric nonalcoholic fatty liver disease. Hepatology . https://doi.org/10.1002/hep.29596
25. Bonekamp S, Tang A, Mashhood A, et al. (2014) Spatial distribution of MRI-determined hepatic proton density fat fraction in adults with nonalcoholic fatty liver disease. J Magn Reson Imaging 39(6):1525–1532

26. Sofue K, Mileto A, Dale BM, Zhong X, Bashir MR (2015) Interexamination repeatability and spatial heterogeneity of liver iron and fat quantification using MRI-based multistep adaptive fitting algorithm. J Magn Reson Imaging 42(5):1281–1290
27. Arulanandan A, Ang B, Bettencourt R, et al (2015) Association between quantity of liver fat and cardiovascular risk in patients with nonalcoholic fatty liver disease independent of nonalcoholic steatohepatitis. Clin Gastroenterol Hepatol 13(8):1513–1520 e1511
28. Vu KN, Gilbert G, Chalut M, et al. (2016) MRI-determined liver proton density fat fraction, with MRS validation: comparison of regions of interest sampling methods in patients with type 2 diabetes. J Magn Reson Imaging 43(5):1090–1099
29. Noureddin M, Lam J, Peterson MR, et al. (2013) Utility of magnetic resonance imaging versus histology for quantifying changes in liver fat in nonalcoholic fatty liver disease trials. Hepatology 58(6):1930–1940
30. Serai SD, Dillman JR, Trout AT (2017) Proton density fat fraction measurements at 1.5- and 3-T hepatic MR imaging: same-day agreement among readers and across two imager manufacturers. Radiology 284(1):244–254
31. Campo CA, Hernando D, Schubert T, et al. (2017) Standardized approach for ROI-based measurements of proton density fat fraction and R2* in the liver. Am J Roentgenol . https://doi.org/10.2214/AJR.17.17812
32. Johnson BL, Schroeder ME, Wolfson T, et al. (2014) Effect of flip angle on the accuracy and repeatability of hepatic proton density fat fraction estimation by complex data-based, T1-independent, T2*-corrected, spectrum-modeled MRI. J Magn Reson Imaging 39(2):440–447
33. Hines CD, Frydrychowicz A, Hamilton G, et al. (2011) T(1) independent, T(2) (*) corrected chemical shift based fat–water separation with multi-peak fat spectral modeling is an accurate and precise measure of hepatic steatosis. J Magn Reson Imaging 33(4):873–881
34. Merriman RB, Ferrell LD, Patti MG, et al. (2006) Correlation of paired liver biopsies in morbidly obese patients with suspected nonalcoholic fatty liver disease. Hepatology 44(4):874–880
35. Ratziu V, Charlotte F, Heurtier A, et al. (2005) Sampling variability of liver biopsy in nonalcoholic fatty liver disease. Gastroenterology 128(7):1898–1906
36. Juluri R, Vuppalanchi R, Olson J, et al. (2011) Generalizability of the nonalcoholic steatohepatitis Clinical Research Network histologic scoring system for nonalcoholic fatty liver disease. J Clin Gastroenterol 45(1):55–58
37. Pournik O, Alavian SM, Ghalichi L, et al. (2014) Inter-observer and intra-observer agreement in pathological evaluation of non-alcoholic fatty liver disease suspected liver biopsies. Hepat Mon 14(1):e15167
38. Strauss S, Gavish E, Gottlieb P, Katsnelson L (2007) Interobserver and intraobserver variability in the sonographic assessment of fatty liver. Am J Roentgenol 189(6):W320–W323
39. Manning PM, Hamilton G, Wang K, et al. (2017) Agreement between region-of-interest- and parametric map-based hepatic proton density fat fraction estimation in adults with chronic liver disease. Abdom Radiol (NY) 42(3):833–841
40. Hong CW, Wolfson T, Sy EZ, et al. (2017) Optimization of region-of-interest sampling strategies for hepatic MRI proton density fat fraction quantification. J Magn Reson Imaging . https://doi.org/10.1002/jmri.25843
41. Haufe WM, Wolfson T, Hooker CA, et al. (2017) Accuracy of PDFF estimation by magnitude-based and complex-based MRI in children with MR spectroscopy as a reference. J Magn Reson Imaging . https://doi.org/10.1002/jmri.25699