Journal of Magnetic Resonance Imaging

RESEARCH ARTICLE OPEN ACCESS

Multi-Center, Multi-Vendor Validation of Simultaneous MRI-Based Proton Density Fat Fraction and R2* Mapping Using a Combined Proton Density Fat Fraction-R2* Phantom

Jitka Starekova[1] | David Rutkowski[2] | Won C. Bae[3] | Hung Do[4] | Ananth J. Madhuranthakam[5] | Vadim Malis[3] | Sheng Qing Lin[5] | Suraj Serai[6] | Takeshi Yokoo[5] | Scott B. Reeder[1,7,8,9,10] | Jean H. Brittain[2] | Diego Hernando[1,8]

[1]Department of Radiology, University of Wisconsin-Madison, Madison, Wisconsin, USA | [2]Calimetrix, Madison, Wisconsin, USA | [3]Department of Radiology, University of California-San Diego, La Jolla, California, USA | [4]Canon Medical Systems, Tustin, California, USA | [5]Department of Radiology, University of Texas Southwestern, Dallas, Texas, USA | [6]Department of Radiology, Children's Hospital of Philadelphia, Philadelphia, Pennsylvania, USA | [7]Department of Biomedical Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA | [8]Department of Medical Physics, University of Wisconsin-Madison, Madison, Wisconsin, USA | [9]Department of Medicine, University of Wisconsin-Madison, Madison, Wisconsin, USA | [10]Department of Emergency Medicine, University of Wisconsin-Madison, Madison, Wisconsin, USA

Correspondence: Diego Hernando (dhernando@wisc.edu)

Received: 19 December 2024 | **Revised:** 14 March 2025 | **Accepted:** 18 March 2025

Funding: NIH R01 EB031886; University of WisconsinMadison Office of the Vice Chancellor for Research and Graduate Education; Wisconsin Alumni Research Foundation; UW-Madison Departments of Radiology and Medical Physics. Phase II SBIR grant: R44EB025729. NIH/NCI R01CA283663 and Cancer Prevention and Research Institute of Texas (CPRIT) RP190049. HD is an employee of Canon Medical Systems USA. SBR, JHB and DH are co-founders and DR employee of Calimetrix.

Keywords: CSE-MRI | PDFF | phantom | quantification | R2* | reproducibility

ABSTRACT

Background: Fat and iron deposition confound measurements of R2* and proton density fat fraction (PDFF), respectively, yet their combined impact on reproducibility is poorly understood.

Purpose: To evaluate the multi-center, multi-vendor reproducibility of PDFF and R2* quantification using a PDFF-R2* phantom.

Study Type: Prospective multi-center, phantom study.

Phantom: Commercial PDFF-R2* phantom with simultaneously controlled combination of PDFF (0%–30%) and R2* (50–600 s^{-1}) values.

Field Strength/Sequence: 1.5-T and 3-T, three-dimensional (3D) multi-echo, spoiled-gradient-echo sequences, in four different centers, each with a different vendor.

Assessment: Two acquisition protocols were used, optimized for moderate R2* (Protocol 1) and high R2* (Protocol 2), respectively. The phantom was imaged multiple times at one of the centers to assess its stability.

Statistical Tests: Intraclass correlation coefficient (ICC), linear regression analysis, reproducibility coefficient (RDC) and repeatability coefficient (RC).

Results: Excellent agreement was observed for PDFF measurements between centers, vendors, field strengths, and protocols (ICC = 0.97). Stratified by protocol, excellent agreement was observed, with ICC = 0.96 (RDC = 6.2%) for Protocol 1 and ICC = 0.99 (RDC = 3.8%) for Protocol 2. Increased variability in PDFF measurements was observed with increasing PDFF and

especially with higher R2*. Excellent agreement was observed for R2* between centers, vendors, field strengths, and protocols (ICC = 0.99). Stratified by protocol, strong agreement was observed, with ICC = 0.988 (RDC = 66.7 s^{-1}) for Protocol 1 and ICC = 0.99 (RDC = 57.7 s^{-1}) for Protocol 2. Higher variability in R2* measurements was observed in vials with higher PDFF or R2*. Stability tests demonstrated an ICC = 1.0 for PDFF and R2*, and RC of 0.4% for PDFF and 12 s^{-1} for R2*.

Data Conclusion: Excellent PDFF and R2* reproducibility was observed across centers, vendors, field strengths, and acquisition protocols. Reproducibility decreased slightly with increasing PDFF and R2*, especially for PDFF measurements in vials with high R2*.

Evidence Level: N/A.

Technical Efficacy: Stage 1.

1 | Introduction

Diffuse liver disease is commonly associated with abnormal intracellular deposition of triglycerides and iron [1, 2]. Metabolic dysfunction-associated steatotic liver disease (MASLD; formerly known as non-alcoholic fatty liver disease) is the leading chronic liver disease globally [1, 3]. Specifically, MASLD is characterized by excessive fat accumulation in the liver and affects both adults and children [3]. This progressive liver condition is associated with a range of hepatic and extrahepatic complications [3, 4]. In addition, liver iron overload, caused by chronic blood transfusions, genetic hemochromatosis, or other chronic liver disease, can lead to liver damage and iron overload cardiomyopathy [5]. Further, the concomitant accumulation of both fat and iron in the liver can worsen disease progression and outcomes [6–8].

Reliable quantitative assessment of liver fat and iron content is important for disease management and evaluating new therapies in clinical trials [9–12]. However, if unaccounted for, signal from fat introduces constructive (in-phase) and destructive (opposed-phase) interference patterns with signal from water, leading to bias in the quantification of iron by MRI [13]. Conversely, the presence of iron leads to rapid MR signal decay that can confound MRI methods used for fat quantification [14]. Importantly, patients with iron overload often have concomitant MASLD, which affects approximately 38% of the general population and up to 14% of children and adolescents [3, 4]. Thus, there is a major clinical need to evaluate patients with concomitant liver fat and iron overload.

Confounder-corrected chemical-shift-encoded MRI (CSE-MRI) enables estimation of proton density fat fraction (PDFF) and R2* as quantitative biomarkers for liver fat and iron content, respectively [4, 12]. While CSE-MRI has demonstrated high accuracy and reproducibility across systems, field strengths, and manufacturers [12, 15, 16], it is important to note that past studies primarily focused on individual biomarkers (PDFF or R2*), but not their simultaneous quantification. This knowledge gap is relevant because the presence of iron and fat can confound CSE-MRI-based measurements of PDFF and R2*, respectively [13, 14]. The effect of these potential errors is non-negligible, as biased measurements of liver fat and iron could impact the diagnosis, staging, and treatment monitoring in patients with concomitant liver fat and iron overload.

Phantoms provide a controlled environment that models relevant parameters, and using phantoms with simultaneously controlled varying levels of fat and iron may help address this knowledge gap [17, 18]. Additionally, testing on multiple platforms and at different centers is needed to ensure generalizability. In this context, PDFF phantoms have been previously used in several multi-center studies [19–21]. However, there is a notable lack of multi-center studies using phantoms that incorporate and simultaneously modulate both PDFF and R2* [22].

Therefore, the purpose of this study was to evaluate the multi-center, multi-vendor reproducibility of CSE-MRI-based simultaneous PDFF and R2* mapping at both 1.5-T and 3-T MRI using a combined quantitative PDFF-R2* phantom.

2 | Materials and Methods

This prospective phantom study was conducted at four participating centers between February 2022 and December 2022.

2.1 | PDFF-R2* Phantom

A single commercial phantom (Model 725 PDFF-R2* Phantom; Calimetrix LLC, Madison, WI, USA) was used. The phantom includes 16 cylindrical vials (outer diameter: 27 mm, length: 60 mm), arranged in an asymmetrical grid of PDFF-R2* values (Figure 1a,b). Each vial contains a unique and simultaneously controlled combination of PDFF values ranging from 0% to 30% and R2* values ranging from 50 to 600 s^{-1} (Figure 1c). The PDFF and R2* values of the vials were selected to span the relevant biological ranges of PDFF and R2* in the human liver [4, 23, 24].

Reference PDFF and R2* values of each vial were verified with measurements from a temperature-corrected, confounder-corrected reconstruction of CSE-MRI data, following standard Calimetrix Quality Management System (QMS) procedures (Calimetrix LLC, Madison, WI, USA).

2.2 | Imaging Experiments

The study included four centers equipped with MRI systems from different manufacturers (Figure 1d): GE HealthCare (Center 1), Philips Healthcare (Center 2), Siemens Healthineers (Center 3), and Canon Medical Systems (Center 4). The following MRI systems were included in the study: 1.5-T Signa Artist and 3-T Signa Premier (GE HealthCare, Waukesha, Wisconsin, USA), 1.5-T Avanto and 2.9-T Skyra (Siemens Healthineers, Erlangen, Bavaria, Germany), 1.5-T Ingenia and 3-T Ingenia (Philips Healthcare, Amsterdam, Netherlands), 1.5-T Vantage Orian and

Plain Language Summary

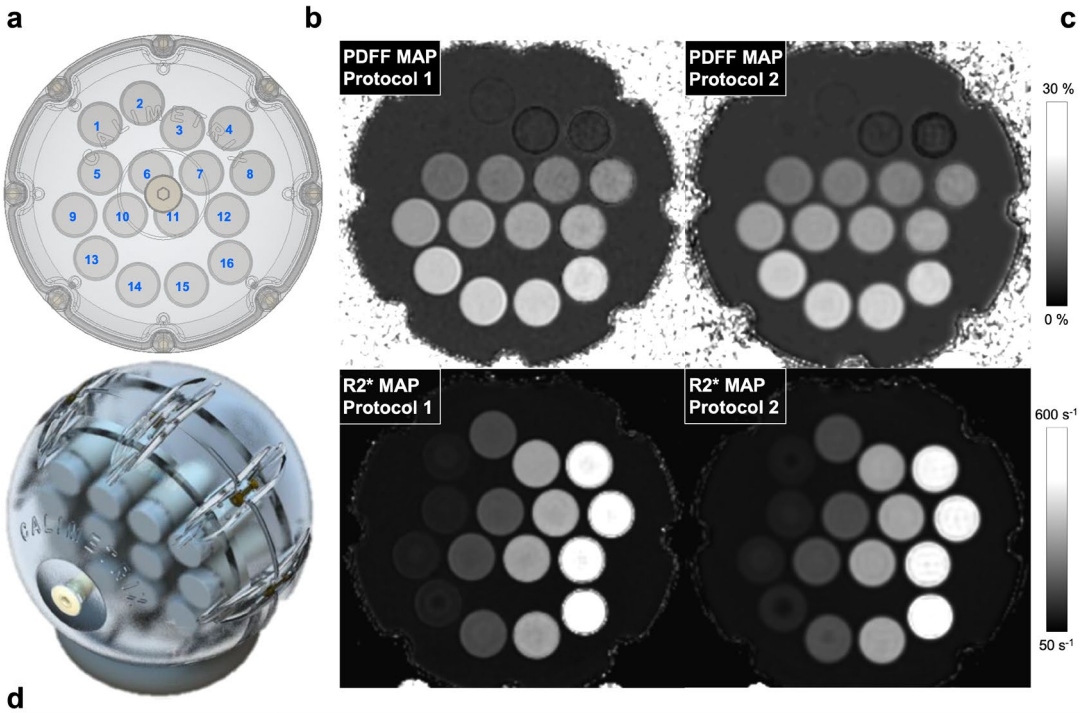
- Measuring fat and iron is essential for diagnosing and managing various liver diseases.
- This study tested how consistently magnetic resonance imaging (MRI) can measure both simultaneously.
- A test object ("phantom"), with known fat and iron concentrations, was imaged at four medical centers using different MRI machines and two different imaging techniques.
- Results showed consistent fat and iron measurements across all machines and techniques.
- This confirms that MRI can measure both fat and iron reliably, even when present together, which is commonly observed in liver disease.
- Using a standardized phantom across multiple sites makes the results more reliable for real-world use.

2.9-T Vantage Galan (Canon Medical Systems, Tochigi, Japan). At each of the four centers, the phantom was imaged at both 1.5-T and 3.0-T (2.89-T for Siemens and Canon), using a three-dimensional (3D) multi-echo, spoiled-gradient-echo (SGRE) sequence with acquisition parameters approximately matched across centers and vendors.

Centers were requested to follow parameters for two different standardized CSE-MRI acquisition protocols. The first protocol was optimized for PDFF quantification in the presence of moderate R2* (protocol 1): echo time $(TE)_1 = 1.0–1.2$ ms for both 1.5-T and 3-T MRI systems; $\Delta TE = 1.8–2.1$ ms for 1.5-T and 0.8–1.0 ms for 3-T MRI systems; 6 echoes obtained in a single echo train at 1.5-T and two interleaved echo trains at 3-T, slice thickness = 4 mm, and flip angle = 3° for both 1.5-T and 3-T MRI systems.

The second protocol was optimized for PDFF and R2* quantification in the presence of high R2* (protocol 2): $TE_1 = 1.0–1.2$ ms for 1.5-T and 0.8–1.1 ms for 3-T; $\Delta TE = 0.7–0.9$ ms (1.5-T) and 0.6–0.8 ms (3-T); 8 echoes obtained in two interleaved echo trains, slice thickness = 4 mm, flip angle = 4° for both 1.5-T and 3-T MRI systems. Tables 1 and 2 summarize the recommended and applied typical imaging parameters for both protocols on each MRI system. These parameters were matched as closely as possible between centers, although exact alignment was not always feasible due to differences in hardware and pulse sequence limitations across MRI systems and vendors.

Protocol 1 is expected to provide relatively high resolution and good image quality for moderate R2* values. However, this protocol may be inadequate for cases with high R2* due to lower SNR at longer echo times. Therefore, protocol 2 was added to address these limitations and potentially improve performance for high R2*.



Vial	Nominal PDFF(%)/R2* (s^{-1})
1	0/50
2	0/150
3	0/350
4	0/600
5	10/50
6	10/150
7	10/350
8	10/600
9	20/50
10	20/150
11	20/350
12	20/600
13	30/50
14	30/150
15	30/350
16	30/600

Field strength | Vendor	GE HealthCare	Philips Healthcare	Siemens Healthineers	Canon Medical Systems
1.5T	Signa Artist	Ingenia	Magnetom Avanto	Vantage Orian
3T	Signa Premier	Ingenia	Magnetom Skyra*	Vantage Galan*

FIGURE 1 | Graphic illustration and photograph of the PDFF-R2*phantom (a) with corresponding PDFF and R2* maps (b). The phantom was constructed as an array of 16 vials submerged in doped fill solution within spherical acrylic housing. Each vial contained a unique combination of PDFF and R2* values (nominal values) as shown in the table on the left (c). The phantom was imaged at four centers with different MR vendors (d). *2.89T.

TABLE 1 | Imaging parameters for protocol 1 acquisition (fat quantification in the presence of moderate R2*).

Parameter	Recommended	Center 1 GE HealthCare	Center 2 Philips Healthcare	Center 3 Siemens Healthineers	Center 4 Canon Medical Systems
1.5T					
Coil, channels (*n*)	Head Head–neck	Head–neck (21)	Head (15)	Head (8)	Head–neck (16)
FOV (cm^2)	26×26	26×26	26×26	26×26	26×26
Matrix size	148×148	148×148	144×144	128×128	144×144
Slice thickness (mm)	4	4	4	4	4
TR (msec)	Min[a]	23	8.9	12.5	8.8
First TE (msec)	1.0 to 1.2	1.1	1.2	1.11	1.2
TE spacing (msec)	1.8 to 2.1	1.9	1.3	2.0	1.2
No. of echo trains/No. of echoes	1/6	1/6	1/6	1/6	1/6
Pixel BW (Hz)	1351	1351	1578	1085	1302
FA (degrees)	3	3	5	3	3
3T[c]					
Coil, channels (*n*)	Head or Head and Neck	Head and Neck (19)	Head (32)	Head (8)	Head and Neck (16)
FOV (cm^2)	26×26	26×26	26×26	26×26	26×26
Matrix size	140–160×140–160	160×160	164×163	128×128	144×144
Slice thickness (mm)	4	4	4	4	4
TR (msec)	Min[a]	7.3	7.5	12.5	7.6
First TE (msec)	1.0–1.2	1.1	1.2	1.11	1.2
TE spacing (msec)	0.8–1.0	0.9	1.0	2.0	1.0
No. of echo trains/No. of echoes	2/6	2/6	1/6	1/6[b]	1/6
Pixel BW (Hz)	1136–1429	1136	1494	1090	1302
FA (degrees)	3	3	3	3	3

Abbreviations: BW, bandwidth; FA, flip angle; FOV, field of view; No, number; TE, echo time; TR, repetition time.
[a]Minimum achievable with the provided echo times and other system constraints.
[b]Bipolar readout.
[c]2.89T for Siemens and Canon.

The phantom was shipped between participating centers (Figure 2) using an overnight courier service within a protective, foam-padded case to minimize any risk of phantom damage during shipping. No chemical heating packs were used, as shipping occurred during warmer months. Temperature indicators inside the phantom case confirmed that the temperature was always maintained above 0°C and below 40°C. When not in use, the phantom was stored at room temperature.

Standardized instructions were provided to all centers. Instructions included allowing the phantoms to equilibrate to room temperature within the MRI environment for at least 30 min prior to data acquisition for optimal image quality and reproducibility. Subsequently, the phantom was placed on the MRI table with the vials aligned parallel to the main magnetic field. To simplify and maximize the reproducibility of phantom positioning across sites, a head or head–neck coil was used for all imaging (Tables 1 and 2). The surface temperature of the phantom was recorded immediately prior to imaging using a sticker temperature sensor adhered to the phantom. Temperature monitoring was performed, as changes in temperature can lead to fat quantification errors as large as 20%, although errors are generally smaller within a temperature range of 15°C–40°C when using complex fitting implemented in vendor reconstructions [25].

To evaluate the integrity of the phantom as well as any potential drift in PDFF and R2* values, the phantom was imaged multiple times at Center 1 at 3-T (Signa Premier; GE HealthCare, Waukesha, WI, USA; software version RX29.1) throughout the study. Two exams were performed during the initial session

TABLE 2 | Imaging parameters for protocol 2 acquisition (fat-iron quantification in the presence of high R2*).

Parameter	Recommended	Center 1 GE HealthCare	Center 2 Philips Healthcare	Center 3 Siemens Healthineers	Center 4 Canon Medical Systems
1.5T					
Coil, channels (*n*)	Head Head–neck	Head–neck (21)	Head (15)	Head (8)	Head–neck (16)
FOV (cm^2)	26×26	26×26	26×26	26×26	26×26
Matrix size	128×128	128×128	132×130	128×128	128×128
Slice thickness (mm)	4	4	4	4	4
TR (msec)	Min[a]	12.6	8.7	11.0	7.8
First TE (msec)	1.0 to 1.2	1.1	1.2	1.11	0.9
TE spacing (msec)	0.7 to 0.9	0.9	0.9	1.2	0.9
No. of echo trains/No. of echoes	2/8	2/8	1/8	1/8[b]	1/6
Pixel BW (Hz)	1302	1302	1706	1090	1302
FA (degrees)	4	4	5	4	4
3T[c]					
Coil, channels (*n*)	Head or Head and Neck	Head and Neck (21)	Head (15)	Head (8)	Head and Neck (16)
FOV (cm^2)	26×26	26×26	26×26	26×26	26×26
Matrix size	100×100	100×100	100×100	128×128	96×96
Slice thickness (mm)	4	4	4	4	4
TR (msec)	Min[a]	6.8	7.9	11.3	7.6
First TE (msec)	0.8–1.1	0.8	1.1	1.1	1.2
TE spacing (msec)	0.6–0.8	0.7	1.1	0.9	1.0
No. of echo trains/No. of echoes	2/8	2/8	1/8	1/8	1/6
Pixel BW (Hz)	1166–2222	2222	2193	1560	1302
FA (degrees)	4	3	4	4	3

Abbreviations: BW, bandwidth; FA, flip angle: No, number; FOV, field of view; TE, echo time; TR, repetition time.
[a]Minimum achievable with the provided echo times and other system constraints.
[b]Bipolar readout.
[c]2.89T for Siemens and Canon.

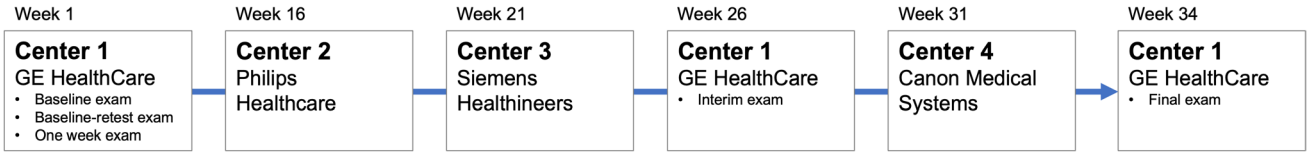


FIGURE 2 | Workflow of phantom shipment and imaging schedule across centers. Imaging at Center 1 was conducted at multiple points throughout the study to assess phantom integrity and potential drift in PDFF and R2* values. Two exams were performed during the initial session (same day): A baseline exam and a retest after repositioning the phantom and reconnecting the coil. Follow-up exams were conducted at 1 week, 6 months (interim exam, after the phantom was shipped from Center 3 to Center 1) and 9 months (final exam, following shipment from Center 4 to Center 1). The phantom was shipped between participating centers using an overnight courier service within a protective, foam-padded case.

(same day): a baseline exam and a retest after repositioning the phantom and reconnecting the coil (Figure 2). Follow-up exams were conducted at 1 week, 6 months (interim exam) and 9 months (final exam, Figure 2). To assess the stability of PDFF values, *protocol 1* was used. To assess the stability of PDFF and R2* values, *protocol 2* was used.

TABLE 3 | Reproducibility coefficient for protocols 1 and 2.

Protocol 1 PDFF overall RDC: 6.2%					**Protocol 2 PDFF overall RDC: 3.8%**				
Vial	**PDFF %**	**R2* s^{-1}**	**SD**	**RDC %**	**Vial**	**PDFF %**	**R2* s^{-1}**	**SD**	**RDC %**
1	0	50	0.33	0.92	1	0	50	0.20	0.57
2	0	150	0.41	1.13	2	0	150	0.28	0.78
3	0	350	0.78	2.15	3	0	350	0.60	1.67
4	0	600	2.55	7.06	4	0	600	1.42	3.94
5	10	50	0.62	1.73	5	10	50	0.68	1.88
6	10	150	0.82	2.29	6	10	150	0.65	1.80
7	10	350	1.03	2.85	7	10	350	1.00	2.76
8	10	600	4.09	11.32	8	10	600	3.83	10.63
9	20	50	1.21	3.36	9	20	50	1.07	2.96
10	20	150	1.96	5.43	10	20	150	1.00	2.78
11	20	350	2.81	7.78	11	20	350	0.94	2.59
12	20	600	2.40	6.65	12	20	600	1.18	3.26
13	30	50	1.90	5.28	13	30	50	1.43	3.95
14	30	150	2.84	7.86	14	30	150	1.23	3.41
15	30	350	4.07	11.27	15	30	350	1.21	3.36
16	30	600	2.70	7.49	16	30	600	1.45	4.01
Protocol 1 R2* overall RDC: 66.7 s^{-1}					**Protocol 2 R2* overall RDC: 57.7 s^{-1}**				
Vial	**PDFF %**	**R2* s^{-1}**	**SD**	**RDC s^{-1}**	**Vial**	**PDFF %**	**R2* s^{-1}**	**SD**	**RDC s^{-1}**
1	0	50	1.93	5.36	1	0	50	1.67	4.64
2	0	150	3.50	9.71	2	0	150	3.90	10.80
3	0	350	11.86	32.88	3	0	350	9.27	25.69
4	0	600	32.36	89.70	4	0	600	22.60	62.65
5	10	50	1.89	5.25	5	10	50	2.65	7.34
6	10	150	6.09	16.88	6	10	150	6.09	16.87
7	10	350	18.17	50.37	7	10	350	14.10	39.08
8	10	600	43.97	121.87	8	10	600	30.84	85.49
9	20	50	2.62	7.27	9	20	50	4.53	12.55
10	20	150	4.49	12.45	10	20	150	7.15	19.82
11	20	350	15.23	42.23	11	20	350	14.35	39.77
12	20	600	42.29	117.23	12	20	600	36.58	101.38
13	30	50	5.54	15.37	13	30	50	10.85	30.08
14	30	150	6.51	18.05	14	30	150	10.28	28.50
15	30	350	18.02	49.96	15	30	350	17.58	48.72
16	30	600	57.60	159.66	16	30	600	54.57	151.26

Abbreviations: PDFF, proton density fat fraction; RDC, reproducibility coefficient; SD, standard deviation.

2.3 | Image Reconstruction

All PDFF and R2* maps were reconstructed automatically at each of the four centers using the vendor-provided reconstruction. All reconstructed images were transferred to Center 1 for analysis. For each vial, a 1.9 cm diameter circular region of interest (ROI) was manually drawn by one radiologist (JS, with 14 years of experience in MRI) in four central slices of PDFF and R2* maps, using OsiriX (version 14.0.1, Pixmeo, Geneva, Switzerland). The mean voxel values from each of the four slices were recorded and averaged for further analysis.

2.4 | Statistical Analysis

Statistical analysis was performed with R (version 4.1.0., tidyverse version 1.3.1, ggplot2 version 3.3.6, irr version 0.84.1, rstatix version 0.7.0; https://www.r-project.org/). For the multicenter, multi-vendor validation of PDFF-R2* mapping, linear regression analysis, intraclass correlation coefficient (ICC) and reproducibility coefficient (RDC) were calculated [26, 27]. Of note, the R2* modulation in the phantom was designed to be field strength-independent between 1.5-T and 3-T MRI systems. For this reason, analysis of R2* data was performed jointly across field strengths in this study.

The stability of the phantom was assessed using the ICC to determine the correlation between longitudinal acquisitions acquired on a single 3-T MRI system at Center 1. Additionally, the repeatability coefficient (RC) was calculated based on the repeated acquisitions from the initial session and follow-up acquisitions at Center 1.

Outlier detection was performed using the rstatix package (rstatix v0.7.2) in R (https://www.r-project.org/), which implements the interquartile range (IQR) method. Values were classified as outliers if they fell outside 1.5 times the IQR below the first quartile or above the third quartile, and as extreme outliers if they fell outside 3 times the IQR.

3 | Results

The PDFF and R2* data from both protocols were successfully collected from all centers, vendors, and systems. The

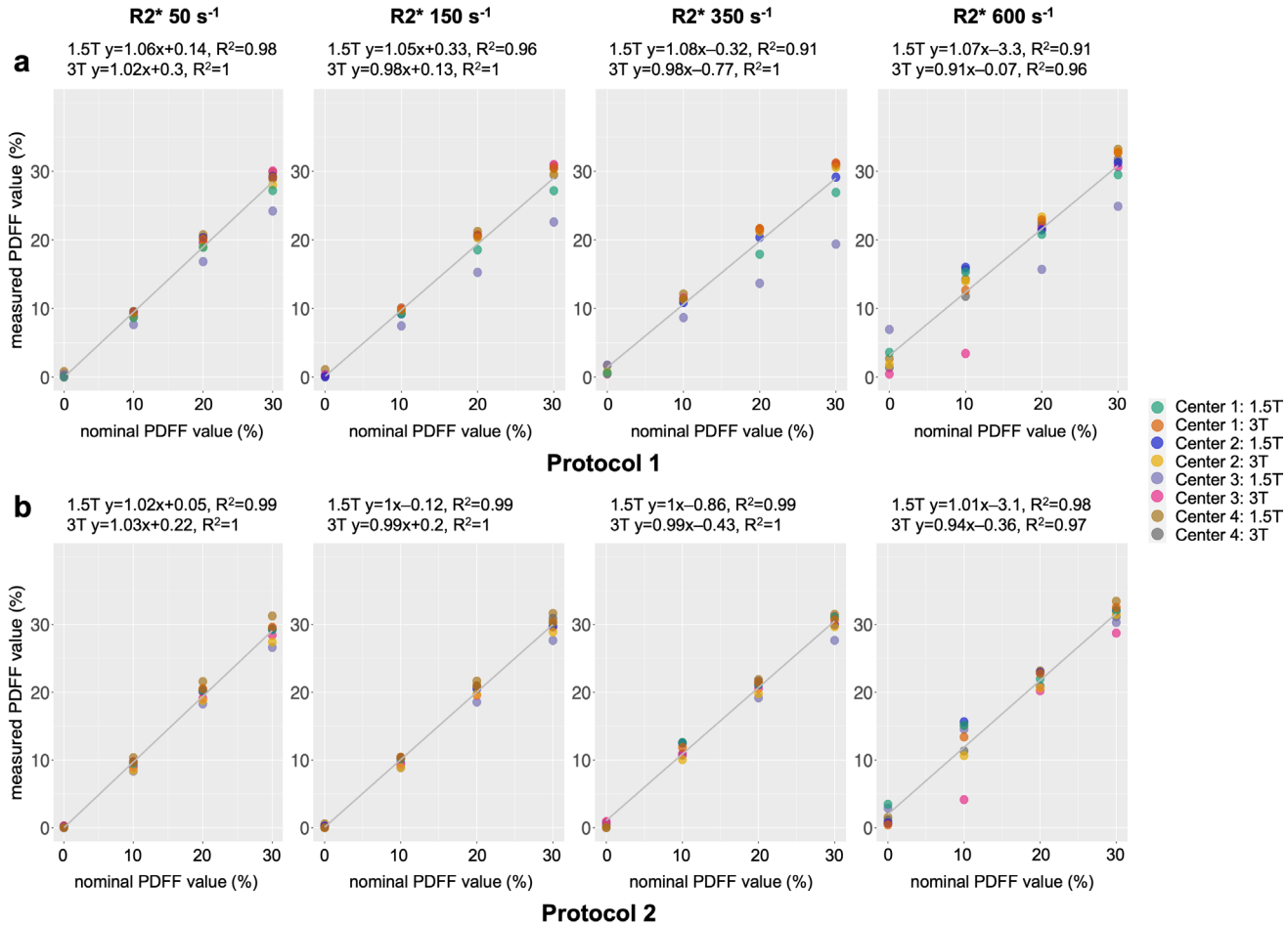


FIGURE 3 | Linear regression shows excellent agreement in PDFF values for low to moderate iron levels across centers and platforms, with higher variability in the presence of high fat and iron content. Depicted are reference- and measured PDFF values from MRI-CSE protocol 1 (a) and protocol 2 (b) across 4 centers and 8 different systems. The columns represent vials with nominal R2* of 50, 150, 350, and 600 s^{-1}, respectively. Vendors: Center 1, GE HealthCare; Center 2, Philips Healthcare; Center 3, Siemens Healthineers; Center 4, Canon Medical Systems. Protocols: Protocol 1 was optimized for PDFF quantification in the presence of moderate R2*, and Protocol 2 was optimized for both PDFF and R2* quantification in the presence of high R2*.

surface temperature of the phantom prior to imaging varied slightly between centers and systems, ranging from 19°C to 23°C.

3.1 | PDFF Reproducibility

Excellent agreement was observed for all PDFF measurements between centers, vendors, field strength, and sequences, with an ICC of 0.97 [95% confidence interval {CI}: 0.95–0.99] and RDC of 5.1%. Stratified per protocol, excellent agreement was observed between centers and systems for *protocol 1*, with an ICC of 0.96 [95% CI: 0.92–0.98], as well as for *protocol 2*, with an ICC of 0.99 [95% CI: 0.97–0.99]. The overall RDC was 6.2% for *protocol 1* and 3.8% for *protocol 2* (Table 3). Variability increased with higher PDFF and R2*, particularly for PDFF measurements in the presence of high R2*, with protocol 2 showing slightly better performance than protocol 1 (Table 3).

Linear regression analysis demonstrated excellent agreement across PDFF values acquired across vendors and field strengths, and between both protocol 1 (R^2 range: 0.91–1.0) and protocol 2 (R^2 range: 0.97–1.0). While some variability of PDFF values was observed at low R2* levels, greater variation was observed in PDFF values with increasing R2* (Figure 3) regardless of protocol (Figure 3). Several values in PDFF measurements for vials with higher iron concentration were lower than expected at Center 3 (Figure 3).

3.2 | R2* Reproducibility

Excellent agreement was observed for all R2* measurements between centers, vendors, field strengths, and protocols with an ICC of 0.99 [95% CI: 0.98–0.996] and RDC of 60.8 s^{-1}. Stratified per protocol, strong agreement was observed between centers, vendors, and field strengths with an ICC of 0.988 [95% CI: 0.976–0.995] for *protocol 1*, and an ICC of 0.99 [95% CI: 0.981–0.996] for *protocol 2*. The overall RDC was 66.7 s^{-1} for *protocol 1* and 57.7 s^{-1} for *protocol 2* (Table 3). Table 3 depicts RDC values using both protocols for all vials. The percentage reproducibility coefficient (% RDC) is given in Table 4.

Linear regression analysis demonstrated excellent agreement for R2* values across centers, vendors, and field strengths for both protocol 1 (R^2 range: 0.99–1.00) and protocol 2 (R^2 range: 0.98–1.00) (Figure 4). Increased variability in R2* values was observed in the vials with higher PDFF and particularly higher R2* (Figure 4).

3.3 | Phantom Stability

The phantom housing and all vials were inspected upon return to Center 1, and no signs of damage were detected. Excellent agreement was observed between initial and follow-up acquisition measurements at Center 1, with an overall ICC of 1.0 [95% CI: 1.0–1.0] for PDFF (Figure 5a) and an ICC of 1.0 [95% CI: 0.999–1.0] for R2* (Figure 5b). From the associated repeatability

TABLE 4 | R2* reproducibility coefficient and percentage reproducibility coefficient for protocol 1 and 2.

	Vial content		Protocol 1 R2* overall RDC: 66.7 s^{-1}			Protocol 2 R2* overall RDC: 57.7 s^{-1}		
Vial No.	PDFF %	R2* s^{-1}	RDC s^{-1}	% RDC	% log RDC	RDC s^{-1}	% RDC	% log RDC
1	0	50	5.36	15.18	3.86	4.64	13.21	3.39
2	0	150	9.71	7.09	1.41	10.80	7.91	1.57
3	0	350	32.88	9.66	1.66	25.69	7.58	1.30
4	0	600	89.70	15.91	2.53	62.65	11.29	1.78
5	10	50	5.25	12.25	3.16	7.34	16.84	4.37
6	10	150	16.88	10.91	2.18	16.87	10.89	2.16
7	10	350	50.37	13.59	2.31	39.08	10.64	1.81
8	10	600	121.87	20.30	3.16	85.49	14.39	2.27
9	20	50	7.27	17.15	4.36	12.55	28.12	7.15
10	20	150	12.45	9.05	1.81	19.82	14.28	2.84
11	20	350	42.23	12.52	2.14	39.77	11.85	2.02
12	20	600	117.23	20.59	3.31	101.38	18.22	2.90
13	30	50	15.37	34.65	8.96	30.08	60.54	14.34
14	30	150	18.05	12.40	2.46	28.50	19.20	3.75
15	30	350	49.96	14.54	2.49	48.72	14.31	2.44
16	30	600	159.66	26.13	4.18	151.26	25.26	4.02

Note: % RDC calculated as RDC/nominal value *100. % log RDC calculated based on log-normalized R2* values as log RDC/log(nominal value)×100.
Abbreviation: RDC, reproducibility coefficient.

acquisitions, the overall RC was 0.4% for PDFF and 12 s^{-1} for R2*. The surface temperature of the phantom prior to imaging remained stable at 19°C for all acquisitions.

4 | Discussion

We performed a phantom study involving 4 centers, 4 vendors, multiple platforms at both 1.5-T and 3-T, and two CSE-MRI protocols to evaluate the reproducibility of CSE-MRI to quantify PDFF and R2*. Through the use of a commercially available phantom with simultaneously modulated PDFF and R2* values, we demonstrated excellent multi-center, multi-vendor reproducibility of fat and iron quantification at both 1.5-T and 3-T MRI systems. Despite excellent reproducibility, we found that reproducibility worsens slightly with increasing PDFF and R2*, most notably for PDFF measurements in the presence of high R2*.

Reproducibility in the context of CSE-MRI for estimation of PDFF and R2* refers to the consistency of measurements across different conditions, such as different MRI platforms, time points, or protocols. For clinical applications, improved reproducibility increases the applicability of consistent thresholds and criteria across different sites, vendors, and time. This consistency facilitates accurate assessment of PDFF and R2* for the diagnosis, treatment efficacy, and safety, and supports robust, generalizable conclusions.

Validation of PDFF and R2* reproducibility across centers, vendors, platforms, and field strengths has been reported, but most studies have been limited exclusively to PDFF or R2* phantoms [21, 28]. Jang et al. conducted a multi-center, multi-vendor, and multi-platform study using a fat-water phantom with varying PDFF values (0%–50%) and commercial CSE-MRI sequences [21]. This study reported a RDC of 10.7% [95% CI: 9.8%–11.6%], which is higher than the RDC found in our study [21]. The potential differences may include phantom construction and temperature stabilization approaches [29].

The reproducibility of PDFF in our study aligns well with previous reproducibility findings from a PDFF phantom and meta-analysis of in vivo PDFF studies across different vendors, field strengths, and reconstruction methods [15, 28]. Our study demonstrated excellent agreement for all PDFF measurements between centers, vendors, field strengths, and protocols high—agreement, in line with previous literature [19]. The several lower-than-expected PDFF values observed at Center 3 in vials with higher iron concentrations may be caused in part by the longer echo spacings used at this Center,

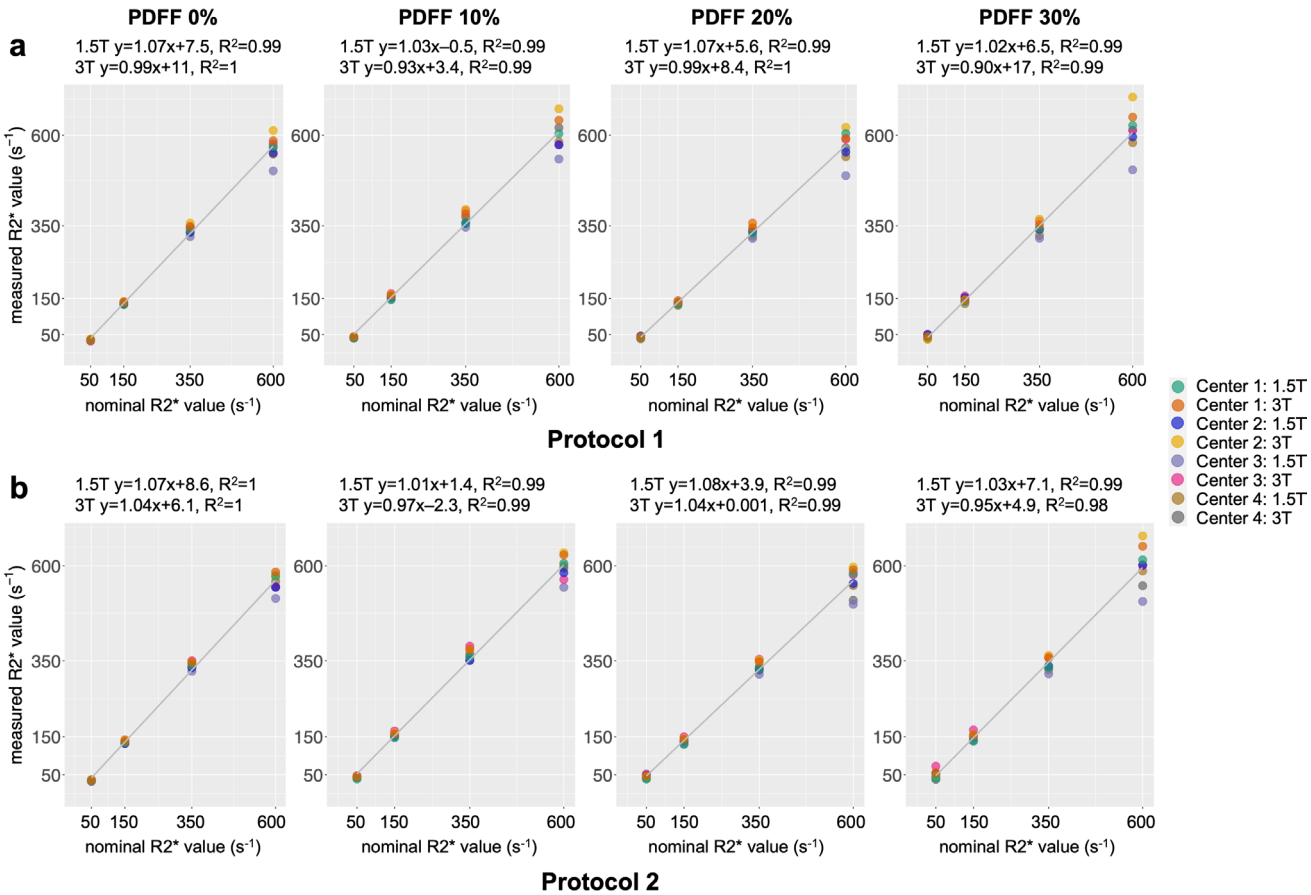


FIGURE 4 | Linear regression shows excellent agreement in R2* values across centers and platforms, with higher variability in the presence of high fat and iron content. Depicted are nominal and measured R2* values from MRI-CSE protocol 1 (a) and protocol 2 (b) across 4 centers and 8 different systems. The columns represent vials with nominal PDFF of 0%, 10%, 20%, and 30%, respectively. Vendors: Center 1, GE HealthCare; Center 2, Philips Healthcare; Center 3, Siemens Healthineers; Center 4, Canon Medical Systems. Protocols: Protocol 1 was optimized for PDFF quantification in the presence of moderate R2*, and Protocol 2 was optimized for both PDFF and R2* quantification in the presence of high R2*.

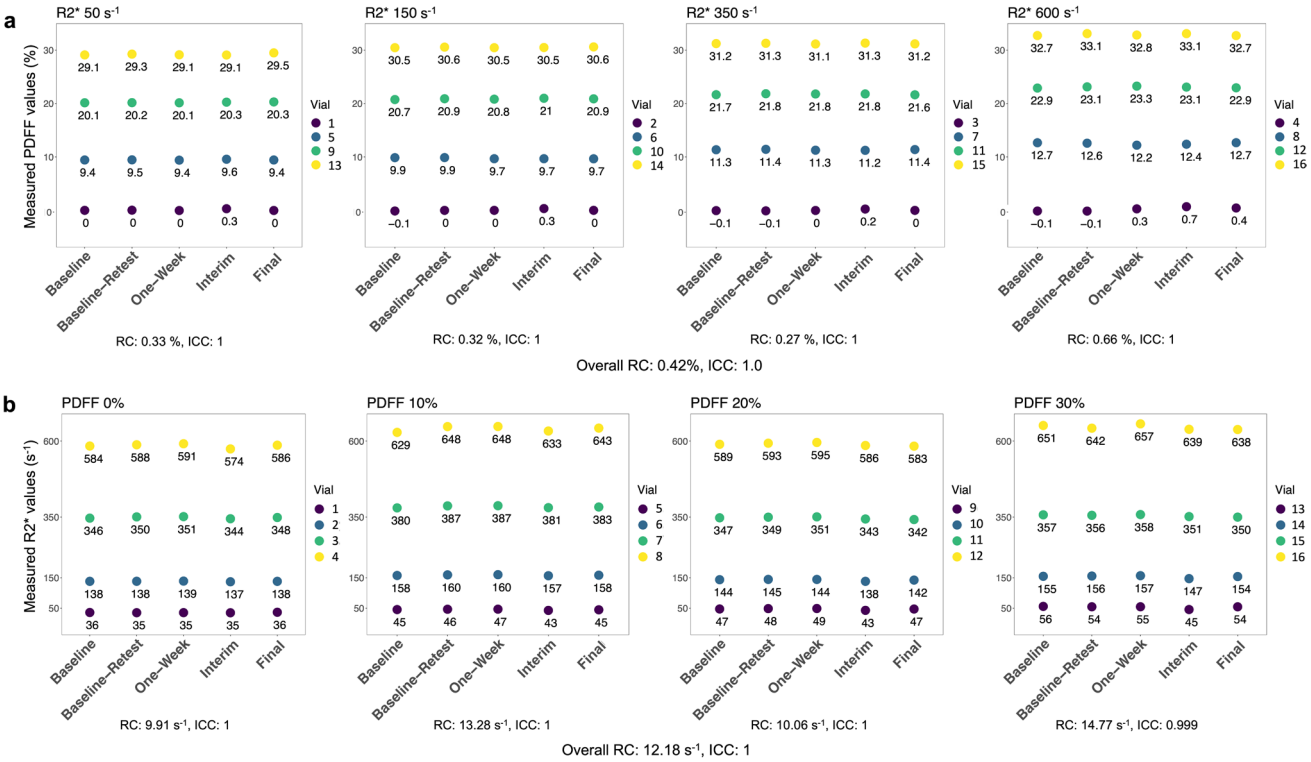


FIGURE 5 | The phantom was imaged at Center 1 several times to test its stability, before and after it was returned to Center 1. Excellent agreement was observed between these acquisitions, with an overall ICC = 1.0 for PDFF and ICC = 1 for R2*. Two exams were performed during the initial session (same day): A baseline exam and a retest after repositioning the phantom and reconnecting the coil. Follow-up exams were conducted at 1 week, 6 months (interim exam) and 9 months later (final exam).

particularly for protocol 1. Longer echo spacings are expected to lead to increased noise amplification and instability at high R2*.

Although there is strong evidence supporting the use of R2*-based liver iron concentration quantification [12, 19], a comprehensive meta-analysis on the reproducibility of R2* is needed, similar to the recent meta-analysis conducted for PDFF [12, 15]. Hernando et al. reported high reproducibility of the calibration between R2* and liver iron concentration in a human study in patients with iron overload [30]. However, a direct comparison with our results is not possible due to differences in study designs and analysis. Our results demonstrated excellent agreement for R2* between centers, vendors, field strengths, and protocols. The RDC increases with higher R2* values, likely due to the increased noise propagation. In contrast, % RDC does not show a monotonic relationship with R2*. This may be due to a combination of high % RDC values at low nominal R2* (since R2* is in the denominator for the calculation of % RDC), and increased noise propagation at high R2*.

With regards to bias, while this study demonstrated excellent reproducibility of CSE-MRI to quantify PDFF and R2*, the maximum acceptable bias required for clinical use remains unclear and is likely application-dependent. However, it is expected that clinical trials and clinical practice will always benefit from the use of quantitative imaging biomarkers (QIBs) with the best possible bias and reproducibility, as demonstrated in previous studies [15, 20, 31]. In the context of clinical trials, understanding the performance of the biomarker would enable well-powered study designs, with the minimum number of subjects needed to demonstrate treatment effects.

Our study advances previous research [21, 28] by demonstrating reproducibility in a controlled setting with a phantom that modulates both PDFF and R2*, simultaneously. As abnormal fat and iron accumulation often coexist, this is an important consideration for patients with both iron overload and concurrent diffuse liver disease, such as MASLD [6]. Simultaneous evaluation of the reproducibility of PDFF and R2* has not been addressed in prior phantom or patient studies. The presence of both fat and iron can cause signal intensity variations due to constructive and destructive interference during gradient-echo acquisitions [12]. Confounder-corrected CSE-MRI addresses this challenge through joint modeling of fat and water signals and R2* decay [4, 32]. However, in the presence of severe iron overload, spectral line broadening leads to merging of fat and water spectral peaks. Based on our data, it may be advantageous to use a CSE-MRI protocol optimized for high R2* in patients with concurrent fat and iron deposition, as it performs slightly better when both PDFF and R2* levels are elevated.

4.1 | Limitations

Although our study included MRI scanners from several leading vendors, it did not include all available manufacturers, such as United Imaging. Vendor-specific online reconstructions for

PDFF and R2* mapping were used in this study. Unlike a unified offline reconstruction approach, online reconstructions can introduce variability that may affect the overall reproducibility. While less controlled, this approach more accurately mimics a real-world clinical setting. Another limitation is the lack of uniform temperature control. While the phantom was temperature-stabilized, maintaining a consistent temperature across sites and systems was not feasible, potentially introducing bias, although no major bias was observed. Variations in the temperature during imaging could partially account for discrepancies between estimated and reference PDFF values. Additionally, the extent of temperature dependency may differ across vendors and reconstruction algorithms [20]. Lastly, our study was conducted solely on phantoms, not ex vivo or in vivo tissue. This approach limits the direct clinical applicability of our results but also allows for more precise control of imaging conditions. While multi-center studies involving 'traveling patients' have been conducted previously [33], such studies present significant logistical challenges compared to phantom studies, where patient transfer is not required. The protocols used in this study can be adopted for in vivo imaging at both 1.5-T and 3-T, but adjustments to factors such as field of view and slice thickness are necessary for clinical applications [14].

5 | Conclusion

Excellent reproducibility of quantitative simultaneous PDFF and R2* measurements may be achieved, as indicated by this multi-center, multi-vendor, multi-platform study using commercial CSE-MRI applications and a commercial phantom with simultaneously modulated PDFF and R2* values.

Acknowledgments

The authors wish to acknowledge support from GE Healthcare and Bracco Diagnostic who provide research support to the University of Wisconsin.

References

1. B. J. Perumpail, M. A. Khan, E. R. Yoo, G. Cholankeril, D. Kim, and A. Ahmed, "Clinical Epidemiology and Disease Burden of Nonalcoholic Fatty Liver Disease," *World Journal of Gastroenterology* 23 (2017): 8263–8276.

2. S. Milic, I. Mikolasevic, L. Orlic, et al., "The Role of Iron and Iron Overload in Chronic Liver Disease," *Medical Science Monitor* 22 (2016): 2144–2151.

3. Z. M. Younossi, M. Kalligeros, and L. Henry, "Epidemiology of Metabolic Dysfunction-Associated Steatotic Liver Disease," *Clinical and Molecular Hepatology* 31 (2024): S32–S50.

4. J. Starekova, D. Hernando, P. J. Pickhardt, and S. B. Reeder, "Quantification of Liver Fat Content With CT and MRI: State of the Art," *Radiology* 301 (2021): 250–262.

5. C. C. Hsu, N. H. Senussi, K. Y. Fertrin, and K. V. Kowdley, "Iron Overload Disorders," *Hepatology Communications* 6 (2022): 1842–1854.

6. J.-P. Kühn, P. Meffert, C. Heske, et al., "Prevalence of Fatty Liver Disease and Hepatic Iron Overload in a Northeastern German Population by Using Quantitative MR Imaging," *Radiology* 284 (2017): 706–716.

7. S. Nishina, M. Korenaga, I. Hidaka, et al., "Hepatitis C Virus Protein and Iron Overload Induce Hepatic Steatosis Through the Unfolded Protein Response in Mice," *Liver International* 30 (2010): 683–692.

8. D. T. Boll, D. Marin, G. M. Redmon, S. I. Zink, and E. M. Merkle, "Pilot Study Assessing Differentiation of Steatosis Hepatis, Hepatic Iron Overload, and Combined Disease Using Two-Point Dixon MRI at 3 T: In Vitro and In Vivo Results of a 2D Decomposition Technique," *American Journal of Roentgenology* 194 (2010): 964–971.

9. S. B. Reeder and J. Starekova, "MRI Proton Density Fat Fraction for Liver Disease Risk Assessment: A Call for Clinical Implementation," *Radiology* 309 (2023): e232552.

10. A. M. Diehl and C. Day, "Cause, Pathogenesis, and Treatment of Nonalcoholic Steatohepatitis," *New England Journal of Medicine* 377 (2017): 2063–2072.

11. J. Patel, R. Bettencourt, J. Cui, et al., "Association of Noninvasive Quantitative Decline in Liver Fat Content on MRI With Histologic Response in Nonalcoholic Steatohepatitis," *Therapeutic Advances in Gastroenterology* 9 (2016): 692–701.

12. S. B. Reeder, T. Yokoo, M. França, et al., "Quantification of Liver Iron Overload With MRI: Review and Guidelines From the ESGAR and SAR," *Radiology* 307 (2023): e221856.

13. D. Hernando, J. H. Kramer, and S. B. Reeder, "Multipeak Fat-Corrected Complex R2* Relaxometry: Theory, Optimization, and Clinical Validation," *Magnetic Resonance in Medicine* 70 (2013): 1319–1331.

14. T. J. Colgan, R. Zhao, N. T. Roberts, D. Hernando, and S. B. Reeder, "Limits of Fat Quantification in the Presence of Iron Overload," *Journal of Magnetic Resonance Imaging* 54 (2021): 1166–1174.

15. T. Yokoo, S. D. Serai, A. Pirasteh, et al., "Linearity, Bias, and Precision of Hepatic Proton Density Fat Fraction Measurements by Using MR Imaging: A Meta-Analysis," *Radiology* 286 (2018): 486–498.

16. D. Hernando, R. J. Cook, N. Qazi, C. A. Longhurst, C. A. Diamond, and S. B. Reeder, "Complex Confounder-Corrected R2* Mapping for Liver Iron Quantification With MRI," *European Radiology* 31 (2021): 264–275.

17. K. E. Keenan, K. V. Jordanova, S. E. Ogier, et al., "Phantoms for Quantitative Body MRI: A Review and Discussion of the Phantom Value," *Magma* 37 (2024): 535–549.

18. R. Zhao, G. Hamilton, J. H. Brittain, S. B. Reeder, and D. Hernando, "Design and Evaluation of Quantitative MRI Phantoms to Mimic the Simultaneous Presence of Fat, Iron, and Fibrosis in the Liver," *Magnetic Resonance in Medicine* 85 (2021): 734–747.

19. D. Hernando, S. D. Sharma, M. Aliyari Ghasabeh, et al., "Multisite, Multivendor Validation of the Accuracy and Reproducibility of Proton-Density Fat-Fraction Quantification at 1.5T and 3T Using a Fat-Water Phantom: Proton-Density Fat-Fraction Quantification at 1.5T and 3T," *Magnetic Resonance in Medicine* 77, no. 4 (2017): 1516–1524, https://doi.org/10.1002/mrm.26228.

20. H. H. Hu, T. Yokoo, M. R. Bashir, et al., "Linearity and Bias of Proton Density Fat Fraction as a Quantitative Imaging Biomarker: A Multicenter, Multiplatform, Multivendor Phantom Study," *Radiology* 298, no. 3 (2021): 640–651, https://doi.org/10.1148/radiol.2021202912.

21. J. K. Jang, S. S. Lee, B. Kim, et al., "Agreement and Reproducibility of Proton Density Fat Fraction Measurements Using Commercial MR Sequences Across Different Platforms: A Multivendor, Multi-Institutional Phantom Experiment," *Investigative Radiology* 54, no. 8 (2019): 517–523, https://doi.org/10.1097/RLI.0000000000000561.

22. S. D. Sharma, D. Hernando, T. Yokoo, et al., "Development and Multi-Center Validation of a Novel Water-Fat-Iron Phantom for Joint Fat and Iron Quantification," Abstract 3274. Annual Meeting of the International Society for Magnetic Resonance in Medicine (ISMRM) Singapore, Published online 2016.

23. M. Obrzut, V. Atamaniuk, K. J. Glaser, et al., "Value of Liver Iron Concentration in Healthy Volunteers Assessed by MRI," *Scientific Reports* 10, no. 1 (2020): 17887, https://doi.org/10.1038/s41598-020-74968-z.

24. S. Verlhac, M. Morel, F. Bernaudin, S. Béchet, C. Jung, and M. Vasile, "Liver Iron Overload Assessment by MRI R2* Relaxometry in Highly Transfused Pediatric Patients: An Agreement and Reproducibility Study," *Diagnostic and Interventional Imaging* 96 (2015): 259–264.

25. D. Hernando, S. D. Sharma, H. Kramer, and S. B. Reeder, "On the Confounding Effect of Temperature on Chemical Shift-Encoded Fat Quantification: Effect of Temperature on CSE Fat Quantification," *Magnetic Resonance in Medicine* 72 (2014): 464–470.

26. M. Gamer, J. Lemon, and IFPS, "irr: Various Coefficients of Interrater Reliability and Agreement. R Package Version 0.84.1," (2019), https://CRAN.R-project.org/package=irr.

27. Radiological Society of North America (RSNA) QIBA, "MRI Technical Performance Indices (RSNA QIBA Technical Report)," (2013) Radiological Society of North America, accessed February 13, 2025, https://qibawiki.rsna.org/images/8/8c/FMRITechnicalPerformanceIndices041613.pdf.

28. E. Schneider, E. M. Remer, N. A. Obuchowski, C. A. McKenzie, X. Ding, and S. D. Navaneethan, "Long-Term Inter-Platform Reproducibility, Bias, and Linearity of Commercial PDFF MRI Methods for Fat Quantification: A Multi-Center, Multi-Vendor Phantom Study," *European Radiology* 31 (2021): 7566–7574.

29. R. Navaratna, R. Zhao, T. J. Colgan, et al., "Temperature-Corrected Proton Density Fat Fraction Estimation Using Chemical Shift-Encoded MRI in Phantoms," *Magnetic Resonance in Medicine* 86 (2021): 69–81.

30. D. Hernando, R. Zhao, Q. Yuan, et al., "Multicenter Reproducibility of Liver Iron Quantification With 1.5-T and 3.0-T MRI," *Radiology* 306 (2023): e213256.

31. Quantitative Imaging Biomarkers Alliance (QIBA) Profile, "MRI-Based Proton Density Fat Fraction (PDFF) of the Liver," accessed February 14, 2025, https://doi.org/10.1148/qiba/20240619.

32. D. Hernando, Y. S. Levin, C. B. Sirlin, and S. B. Reeder, "Quantification of Liver Iron With MRI: State of the Art and Remaining Challenges: Liver Iron Quantification Using MRI," *Journal of Magnetic Resonance Imaging* 40 (2014): 1003–1021.

33. T. He, P. Kirk, D. N. Firmin, et al., "Multi-Center Transferability of a Breath-Hold T2 Technique for Myocardial Iron Assessment," *Journal of Cardiovascular Magnetic Resonance* 10 (2008): 11.