PLOS ONE

RESEARCH ARTICLE

Repeatability and reproducibility of multiparametric magnetic resonance imaging of the liver

Velicia Bachtiar[1]*, Matthew D. Kelly[1], Henry R. Wilman[1,2], Jaco Jacobs[1], Rexford Newbould[1], Catherine J. Kelly[1], Michael L. Gyngell[1], Katherine E. Groves[1], Andy McKay[1], Amy H. Herlihy[1], Carolina C. Fernandes[1], Mark Halberstadt[1], Marion Maguire[1], Naomi Jayaratne[1], Sophia Linden[1], Stefan Neubauer[1,3], Rajarshi Banerjee[1]

1 Perspectum Diagnostics Ltd, Oxford, United Kingdom, **2** Department of Life Sciences, University of Westminster, London, United Kingdom, **3** Oxford Centre for Clinical Magnetic Resonance Research, Radcliffe Department of Medicine, University of Oxford, Oxford United Kingdom

* velicia.bachtiar@perspectum-diagnostics.com

OPEN ACCESS

Citation: Bachtiar V, Kelly MD, Wilman HR, Jacobs J, Newbould R, Kelly CJ, et al. (2019) Repeatability and reproducibility of multiparametric magnetic resonance imaging of the liver. PLoS ONE 14(4): e0214921. https://doi.org/10.1371/journal.pone.0214921

Editor: Peter Lundberg, Linköping University, SWEDEN

Received: June 11, 2018

Accepted: March 24, 2019

Published: April 10, 2019

Data Availability Statement: The minimal anonymised data set has been uploaded to the Open Science Framework repository: URL: https://osf.io/68a9g/ DOI: 10.17605/OSF.IO/68A9G.

Funding: The study was funded by Perspectum Diagnostics. The funder provided support in the form of salaries for authors V.B., M.D.K., J.J., R.N., C.J.K., M.L.G., K.E.G., A.M., A.H.H., C.C.F., M.H., M.M., N.J., S.L., and R.B., but did not have any additional role in the study design, data collection and analysis, decision to publish, or preparation of

Abstract

As the burden of liver disease reaches epidemic levels, there is a high unmet medical need to develop robust, accurate and reproducible non-invasive methods to quantify liver tissue characteristics for use in clinical development and ultimately in clinical practice. This prospective cross-sectional study systematically examines the repeatability and reproducibility of iron-corrected T1 (cT1), T2*, and hepatic proton density fat fraction (PDFF) quantification with multiparametric MRI across different field strengths, scanner manufacturers and models. 61 adult participants with mixed liver disease aetiology and those without any history of liver disease underwent multiparametric MRI on combinations of 5 scanner models from two manufacturers (Siemens and Philips) at different field strengths (1.5T and 3T). We report high repeatability and reproducibility across different field strengths, manufacturers, and scanner models in standardized cT1 (repeatability CoV÷ 1.7%, bias -7.5ms, 95% LoA of -53.6 ms to 38.5 ms; reproducibility CoV 3.3%, bias 6.5 ms, 95% LoA of -76.3 to 89.2 ms) and T2* (repeatability CoV÷ 5.5%, bias -0.18 ms, 95% LoA -5.41 to 5.05 ms; reproducibility CoV 6.6%, bias -1.7 ms, 95% LoA -6.61 to 3.15 ms) in human measurements. PDFF repeatability (0.8%) and reproducibility (0.75%) coefficients showed high precision of this metric. Similar precision was observed in phantom measurements. Inspection of the ICC model indicated that most of the variance in cT1 could be accounted for by study participants (ICC = 0.91), with minimal contribution from technical differences. We demonstrate that multiparametric MRI is a non-invasive, repeatable and reproducible method for quantifying liver tissue characteristics across manufacturers (Philips and Siemens) and field strengths (1.5T and 3T).

the manuscript. The specific roles of these authors are articulated in the 'author contributions' section.

Competing interests: Authors are employees of Perspectum Diagnostics. V.B., M.D.K., J.J., R.N., C. J.K., M.L.G., K.E.G., A.M., A.H.H., C.C.F., M.H., M. M., N.J., S.L., and R.B. are employees at Perspectum Diagnostics. V.B., M.D.K., H.R.W., J. J., R.N., C.J.K., M.L.G., A.H.H., C.C.F., M.M., S.L., S.N., and R.B. are shareholders of Perspectum Diagnostics. S.N. and R.B. are founders of Perspectum Diagnostics. This commercial affiliation does not alter our adherence to PLOS ONE policies on sharing data and materials.

Abbreviations: CoV, Coefficient of Variation; cT1, corrected T1; ICC, Intraclass correlation coefficient; LoA, Limits of Agreement; MRI, Magnetic Resonance Imaging; PDFF, Proton Density Fat Fraction.

Introduction

As the burden of non-alcoholic fatty liver disease (NAFLD) reaches epidemic levels in developed countries [1], [2], there is a pressing need to develop non-invasive, standardised, and quantitative methods [3]. Liver biopsy has long been the gold standard for staging liver disease, yet it is painful, prone to sampling variability [4], has poor inter-observer concordance [5] and carries a risk of complications [6]. Magnetic Resonance Imaging (MRI)-based methods are attractive as they are sensitive to subtle differences in tissue composition, can sample the entire liver, and yield objective quantitative measurements that can contribute to prospective patient management [7]–[9].

Multiparametric MRI is a safe and non-invasive method for quantification of liver tissue characteristics. Images for quantification of hepatic fat from proton density fat fraction (PDFF) maps, T2*, and iron-corrected T1 (cT1) can be rapidly obtained during abdominal breath-hold acquisitions without the need for contrast agents or additional external hardware [8], [10]. Iron correction of T1 (cT1) is necessary to address the confounding effects of excess iron, which is common in chronic liver disease. Liver *MultiScan* (LMS, Perspectum Diagnostics, Oxford, UK) is a software application that can be used with supported MR-systems to correct T1 for the effects of excess iron, and thus, to calculate cT1 from T1 and T2* maps, and standardise to a 3T field strength [10]. This method has been shown to have high diagnostic accuracy for the assessment of liver fibrosis compared to histology [8], predict clinical outcomes in patients with mixed liver disease aetiology [7], identify patients with non-alcoholic steatohepatitis (NASH) and cirrhosis [9], reliably excludes clinically significant liver disease with superior negative predictive value (83.3%) to liver stiffness (42.9%) and is cost-effective in diagnosing NAFLD [11], [12]. Additionally, a recent two-centre study showed excellent test-retest reliability for multiparametric MRI derived metrics (CoV range of 1.4% to 2.8% for cT1) in 22 healthy volunteers [13], indicating good technical precision of this method.

The reliability, or precision of metrics are defined as the extent to which measurements can be reproduced under different conditions such as different scanner field strengths, manufacturers, and models (reproducibility), and reflects the degree of agreement between repeated measurements under identical or near-identical conditions (scan-rescan repeatability) [14]. To be clinically useful, metrics also need to be effective in measuring the heterogeneity of physiological and pathological values in the population [15]. The ability to standardise a measurement across different MR scanner field strengths, manufacturers and models is particularly relevant in the context of clinical practice and multi-site clinical trials.

The purpose of this study was to systematically test the repeatability, reproducibility, and intra- and inter-operator reliability of cT1, T2*, and PDFF measurements across scanner field strength, manufacturer, and model in human participants and phantoms. The performance of T1-mapping standardisation was also evaluated in phantoms.

Materials and methods

Study design and population

Sixty-one participants (aged 22–80, mean 42 years; 25 males; BMI 18–39, mean 25) gave their written informed consent to participate. This study received ethical approval from the South Central Oxford Research Ethics Committee C (Ref: 17/SC/0205). Participants included those with mixed liver disease aetiology (n = 32) and those without any history of liver disease (n = 29) in order to represent a wide range of values of hepatic fat, iron, and fibro-inflammatory status. Exclusion criteria included the presence of MRI contraindications and the inability to obtain informed consent. MR operators and data analysts were blinded to the indication of

participants with liver disease and those without. All participants underwent two serial multiparametric MRI examinations per scanner on at least two different scanners in pseudorandomised order (Fig 1). Same scanner scan-rescan were done on the same day and the time between different scanners ranged from same-day up to 1 week. Participants were instructed to take nothing by mouth for 4 hours before their scan time.

Phantom multiparametric MRI

Phantoms were manufactured to span the normal and clinically relevant range of values expected in the liver to reflect the heterogeneity within the population of interest [16]. Three phantoms, each specific to T1, T2*, and PDFF were manufactured. T1 phantoms were agar gel-based using $NiCl_2$ as the paramagnetic relaxation modifier (range: 338-1075ms at 1.5T and 351-1137ms at 3T). T2* phantoms were aqueous solutions of $MnCl_2$ (range: 3-70ms at 1.5T and 2-43ms at 3T). PDFF phantoms were peanut oil and agar gel-based (0–100% at 1.5T and 3T) manufactured according to the methods of Hines and colleagues [16] (Sigma-Aldrich, UK).

Phantoms were scanned on four Siemens (Avanto^{fit} 1.5T, E11C, MyoMaps; Prisma 3T, E11C, MyoMaps; Skyra 3T, E11C, MyoMaps; Siemens Healthineers) and four Philips (Ingenia 1.5T, 5.3.0, CardiacQuant; Ingenia 3T, 5.3.0, CardiacQuant; Achieva 1.5T, R5.3, CardiacQuant; Achieva dStream 1.5T, R5.3, CardiacQuant; Philips Healthcare) scanners. MyoMaps for Siemens systems and CardiacQuant for Philips systems are commercially available modified Look-Locker inversion recovery (MOLLI) T1-mapping sequences [17]. All phantom measurements were performed with a simulated ECG triggering at 1 beat/s. Differences in the MRI sequences used on Siemens and Philips platforms produce systematic differences in fitted

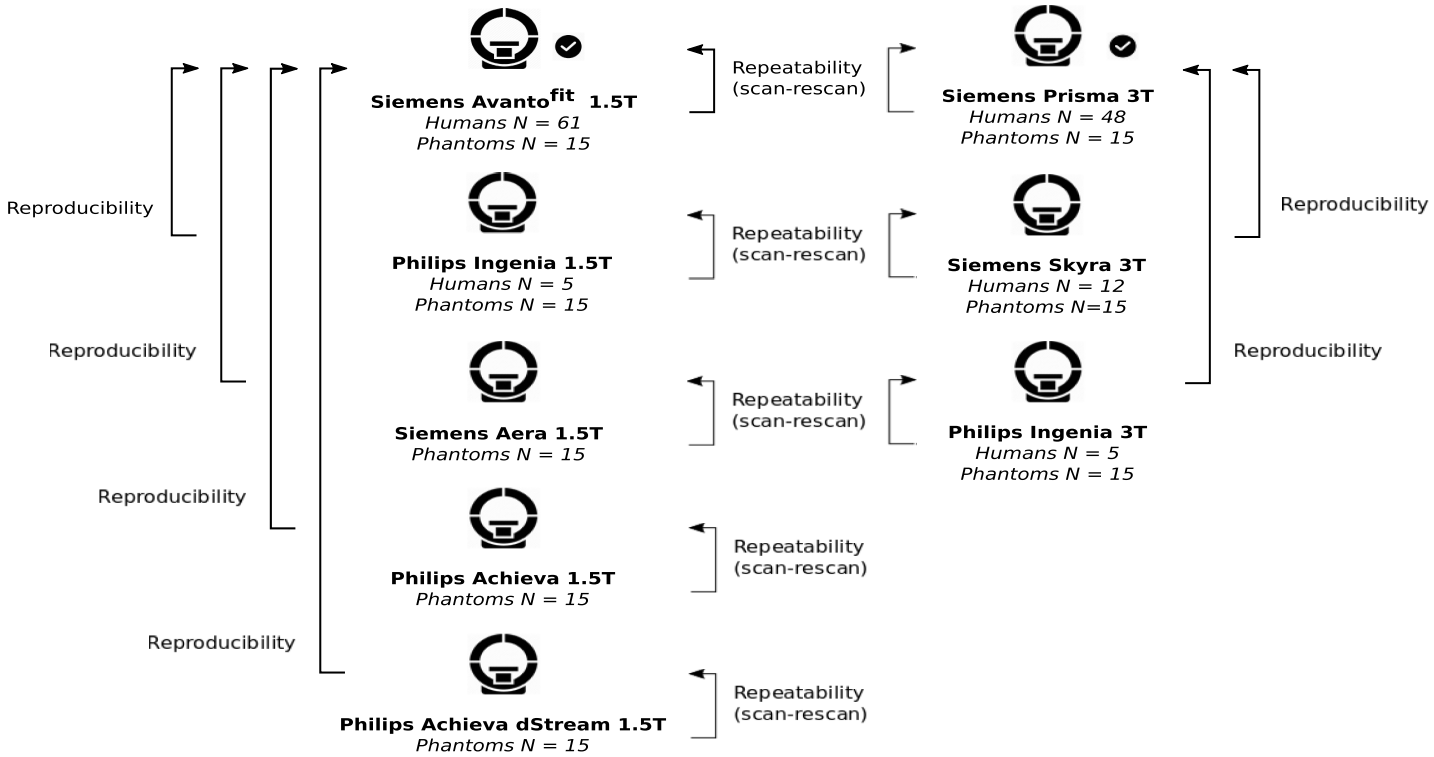

Fig 1. Study design. Two manufacturers (Siemens and Philips) and a range of scanner models were used to systematically test the repeatability and reproducibility of multiparametric-MRI derived measurements in human participants and phantoms.

https://doi.org/10.1371/journal.pone.0214921.g001

T1 values. These quantitative differences were resolved by using distinct, separately acquired phantom measurements to generate linear mapping functions to standardise the values obtained on one system to those from another at the same nominal magnetic field strength (1.5T or 3T). All 3T systems were linearly mapped to the Siemens Prisma 3T, and all 1.5T systems to the Siemens Avanto^{fit} 1.5T, defined as the reference scanners, see Supporting Information S2 File.

Human multiparametric MRI

All human MR scans were performed with participants lying supine on three Siemens (Avanto^{fit} 1.5T, E11C, MyoMaps, OCMR Oxford; Prisma 3T, E11C, MyoMaps, OCMR Oxford; Skyra 3T, E11C, MyoMaps, Southampton General Hospital, UK; Siemens Healthineers) and two Philips (Ingenia 1.5T, 5.3.0, CardiacQuant, Leiden University Medical Centre; Ingenia 3T, 5.3.0, CardiacQuant, Leiden University Medical Centre; Philips Healthcare) scanners. Local radiographers at each imaging centre were trained on the protocol and performed the scans in this study. Single transverse slices were captured through the centre of the liver through the porta hepatis. The individual components of the multiparametric MR protocol consist of T1, T2*, and PDFF-mapping. Full details of the scanning sequences for each scanning platform can be found in Supporting Information S1 File. Linear mappings to reference scanners were performed in the same manner as phantoms, as described above. Any bias introduced by elevated iron was removed from the T1-measurements, yielding the iron-corrected T1 (cT1) as previously described [8], [10]. All human scans on both field strengths used the Siemens Prisma 3T as the reference scanner.

Image processing

Anonymised MR data were analysed off-site using Liver *MultiScan* software (Version 2, Perspectum Diagnostics, UK). Image analysts were trained in abdominal anatomy and images with artefacts were referred to a team of experienced MR physicists for evaluation as previously described [18]. Out of the 138 scans that were completed, 3 scans were unquantifiable due to acquisition errors and in 7 instances due to problems with the scanner cooling system (unrelated to this study and protocol), resulting in a scan success rate of 93%. For each acquisition, three 15mm diameter circular regions of interest (ROIs) were selected on the transverse cT1, T2*, and PDFF maps to cover a representative sample of the liver parenchyma. To assess intra-reader variability, analyst 1 (AN1) re-measured the values for all participants and scan repeats in a randomised order. The time between re-reads was greater than 1 day. To examine inter-reader variability, the first read from analyst 1 (AN1) was compared to an independent read from analyst 2 (AN2). Analysts were blinded to all participant and scanner information.

Statistical analysis

Statistical analyses were carried out using R 3.1.1 [19]. The Bland-Altman method was used to investigate the repeatability and reproducibility between different scanner models against the reference scanners for each metric in phantom (T1, T2*, and PDFF) and human (cT1, T2*, and PDFF) measurements. Repeatability (scan-rescan) was assessed as the closeness of agreement using identical equipment (same scanner field strength, manufacturer, and model). Reproducibility was assessed as the closeness of agreement under varying circumstances (different scanner field strength, manufacturer, and model), such as would be encountered in a multi-centre setting. Limits of Agreement were calculated as the mean of the differences plus and minus 1.96 times the standard deviation of the differences. Repeatability and

reproducibility coefficients are reported as 1.96 times the standard deviation of the differences. Mean coefficients of variation are the mean of the coefficients of variation for each individual.

To further interrogate the reliability of the cT1 metric, a Linear Mixed Effects (LME) approach was implemented using the nlme package [20] in R [19]. LME modelling has been demonstrated to be a superior method to common alternatives such as repeated measures ANOVA or simple paired students t-test as it provides greater statistical power and is robust in the face of missing data [21]. Importantly, LME models for replication data separately and effectively model variance due to within and between subject factors [15], [22]. To assess the variance that could be accounted for by each explanatory variable, the intraclass correlation coefficient (ICC) was calculated to determine the proportion of the total variability in the observations that is due to the differences between pairs.

Results

Standardisation of phantom measurements

We tested the performance of the standardisation of T1 maps across different scanner field strengths, manufacturers, and models using phantom measurements. Bland-Altman analysis of phantom-derived mappings from 90 acquisitions across scanner models and software versions before and after standardization (Fig 2) showed a clear reduction in bias (1.5T: from -23ms to -3.1ms; 3T: from -14ms to -7.8ms), tightening of the 95% Limits of Agreement (LoA) (1.5T: from -66.9ms– 20.4ms, to -24.8ms– 18.6ms; 3T: from -38.1ms– 10.5ms, to -24.8ms– 9.19ms) and a decrease in the mean coefficient of variation (CoV) (1.5T: 1.5% to 0.77%; 3T: 2.9% to 1.1%).

Repeatability and reproducibility of phantom measurements

Standardized T1 from phantom-derived mappings demonstrated high repeatability (CoV 0.16%, bias -0.02 ms, 95% LoA of -4.7 to 4.7 ms) and reproducibility (CoV 1%, bias -4.7 ms, 95% LoA of -25.3 ms to 15.9 ms). $T2^*$-mappings showed good repeatability (CoV 1.1%, bias 0.08 ms, 95% LoA of -0.67 to 0.84 ms) and reproducibility (CoV 3%, bias 0.24ms, 95% LoA of -1.62ms to 2.1ms). Similarly, PDFF measurements also showed good repeatability (CoV 9.7%, bias -0.12%, 95% LoA of -1.4 to 1.14%) and reproducibility (CoV 14%, bias 0.16%, 95% LoA of -4.2% to 4.53%) across different scanner field strengths, manufacturers, and models (Fig 3).

Repeatability and reproducibility of human measurements

Standardized cT1 in participants demonstrated high repeatability (CoV 1.7%, bias -7.5 ms, 95% LoA of -53.6 to 38.5 ms) and reproducibility (CoV 3.3%, bias 6.5 ms, 95% LoA of -76.3 ms to 89.2 ms) across different scanner field strengths, manufacturers, and models. $T2^*$-mappings showed good repeatability (CoV 5.5%, bias -0.18 ms, 95% LoA of -5.4 to 5.1 ms) and reproducibility (CoV 6.6%; bias -1.7ms; 95% LoA of -6.6ms to 3.2 ms). Similarly, PDFF measurements also showed good repeatability (CoV 14%, bias -0.04%, 95% LoA of -0.84 to 0.76%) and reproducibility (CoV 17%, bias 0.06%, 95% LoA of -0.69 to 0.82%) across different scanner field strengths, manufacturers, and models (Fig 4).

To interrogate the cT1 metric further, a random-effects model was generated to determine the variation that could be accounted for by each explanatory variable: scanner type ($Avanto^{fit}$ 1.5T, Prisma 3T, Skyra 3T, Ingenia 1.5T, Ingenia 3T), scan repeat (SR1, SR2), analyst (AN1, AN2), and analysis repeat (AR1, AR2). Inspection of the model indicated that most of the variance in cT1 could be accounted for by study participants (ICC = 0.91), with minimal

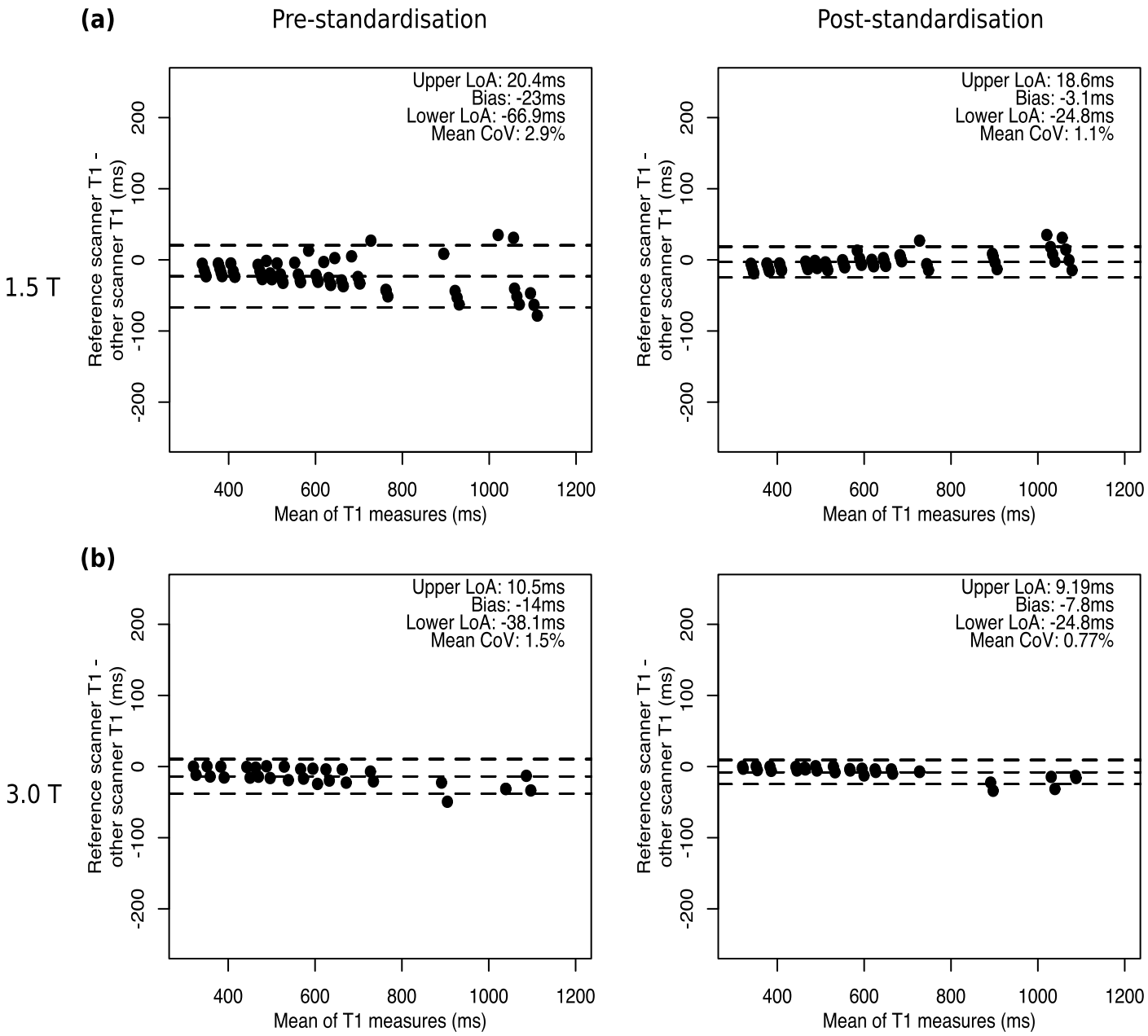

Fig 2. Phantom T1 Standardisation. Bland-Altman plots demonstrating T1 measurements in phantoms before and after standardisation at **(a)** 1.5T and **(b)** 3T.

https://doi.org/10.1371/journal.pone.0214921.g002

contribution from the other explanatory variables (scanner type = 0.04, scan repeat = 0.003, analyst = 0, analysis repeat = 0, residual = 0.05).

Discussion

The primary goal of this study was to systematically test the repeatability and reproducibility of multiparametric-MRI derived measurements across scanner field strength, manufacturer and model in human participants and phantoms. We report the overall repeatability and reproducibility of standardised cT1, T2*, and PDFF measurements.

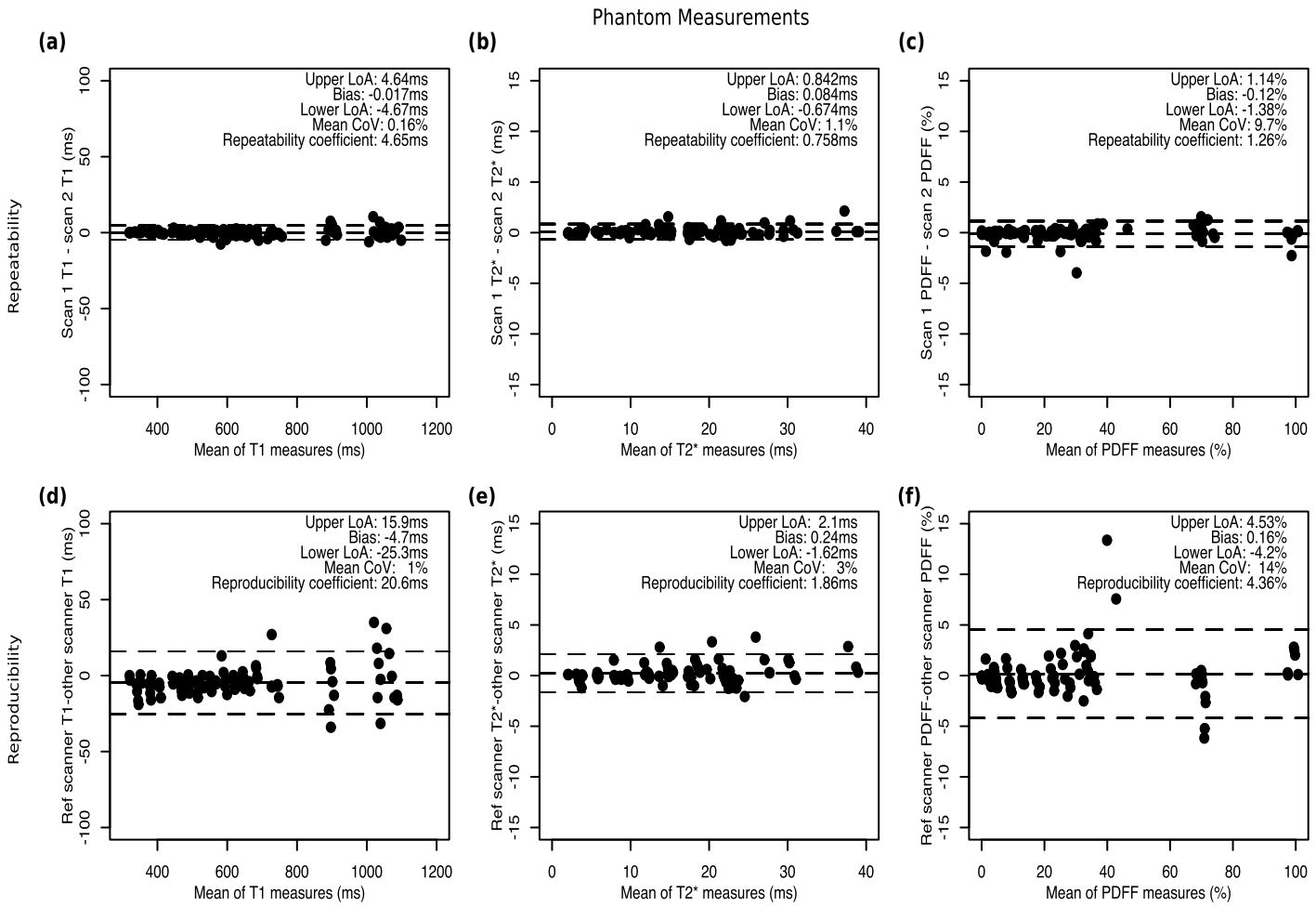

Fig 3. Repeatability and reproducibility of phantom measurements. Bland-Altman plots from phantom measurements across manufacturer and field strength for **(a)** T1, **(b)** T2*, and **(c)** PDFF.

https://doi.org/10.1371/journal.pone.0214921.g003

High repeatability and reproducibility was demonstrated in each metric tested. We report a 3.3% CoV in cT1 measurements across different manufacturer, field strength, and scanner model combinations on 61 participants who had mixed liver disease aetiology as well as those without any history of liver disease to represent the wide range of physiological values in the population. Interrogation of the cT1 metric indicated that most of the variance could be accounted for by study participants (ICC = 0.91), with minimal contributions from scanner type and scan repeat, further supporting the high reproducibility of this measurement.

In a recent study, Harrison and colleagues [23] reported repeatability of cT1, MR Elastography (MRE), and shear-wave ultrasonic elastography (LSM) to reveal CoVs of 3.1%, 11%, and 40% respectively. Similarly, Trout and colleagues [24] reported an average of 10.7% CoV in liver stiffness measurements across different manufacturer, field strength, and sequence combinations on 24 healthy adult volunteers with MRE [24]. However, it is not possible to compare the precision performance of these methods using CoV alone, as the underlying physiological properties and clinically-relevant dynamic range of the techniques are different, and in the Trout and colleagues study, subjects with known liver disease were not included.

Hines et al [25] reported that liver stiffness measurements from MRE varied by 8% between examinations in the same patient performed on the same day, and this increased to 12% when

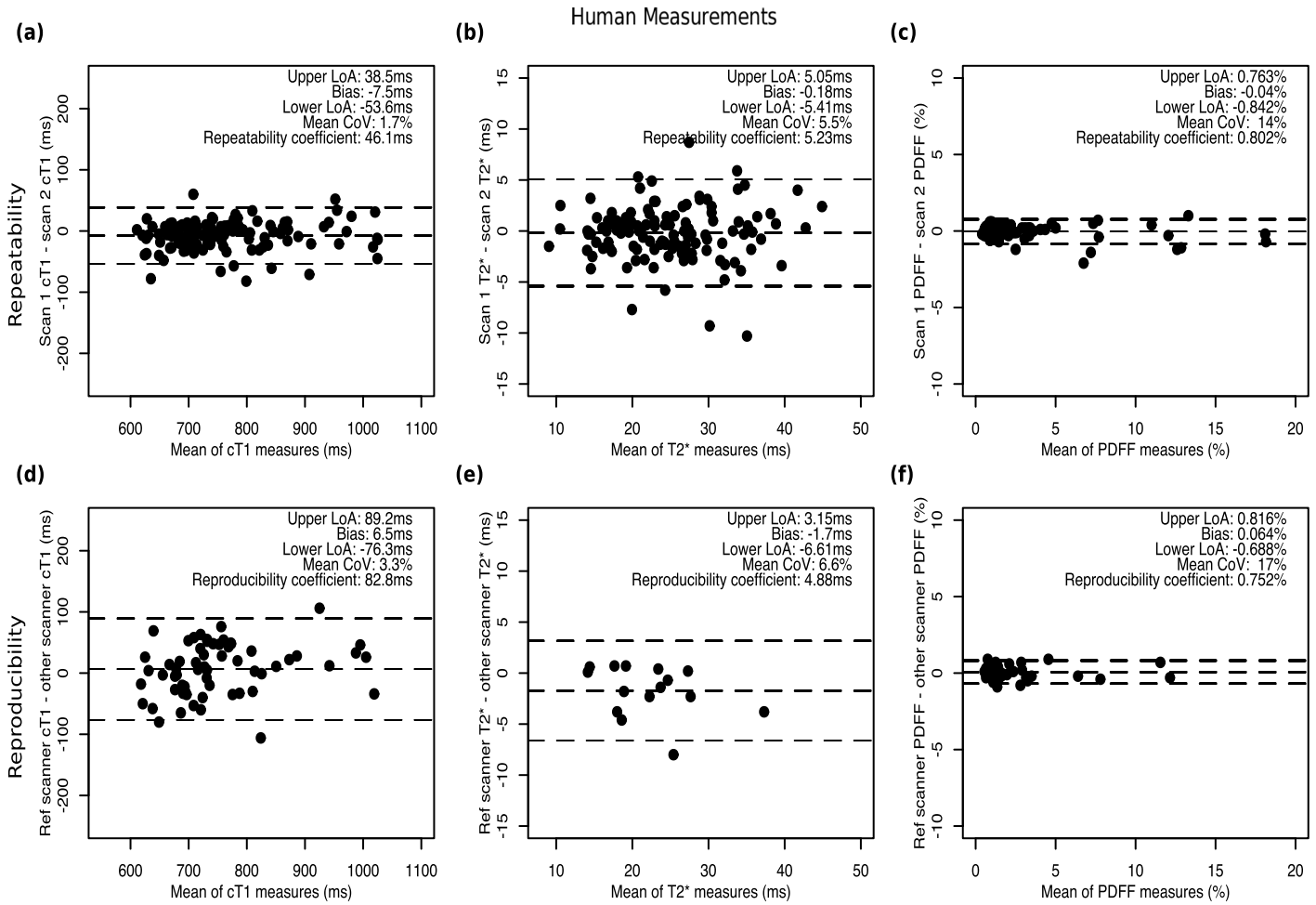

Fig 4. Repeatability and reproducibility of human multiparametric MRI measurements. Bland-Altman plots from human measurements across manufacturer and field strength for **(a)** cT1, **(b)** T2*, and **(c)** PDFF.

https://doi.org/10.1371/journal.pone.0214921.g004

examined on different days separated by 2–4 weeks. In our study same scanner repeatability measurements were performed on the same day and reproducibility on different scanners were performed either on the same day or up to 1-week in between. It is possible that the short time period between serial examinations may have led to an underestimation of physiological variability and consequently a narrower cT1 range within subjects, and it is possible that this may increase with intermediate (e.g. 1 week) and longer (e.g. 6-months) time intervals. Future investigations could define within-subject variability in cT1 measurements to characterise longitudinal fluctuations in this metric.

In a recent study, Bane and colleagues [26] tested T1 repeatability and reproducibility in a T1 phantom across 10 MRI scanners. Using an optimized inversion recovery spin echo technique, they report a median repeatability CoV of 0.3%, and reproducibility CoV of 8.21% at 1.5T and 5.46% at 3T. One site in that study also ran a MOLLI experiment as in this study; the repeatability CoV was reported at 0.68% with a standard error of 4.64%.

PDFF has been recognised as the best current metric for a standardised MR-based biomarker of tissue fat concentration [27]. A meta-analysis of pooled data collected from 28 published studies demonstrated high precision of MR-PDFF across different field strengths,

manufacturers and reconstruction methods, with repeatability and reproducibility coefficients of 2.99% and 4.12% respectively [28]. We report a repeatability coefficient of 0.8% and reproducibility coefficient of 0.75%, indicating excellent precision of this metric, in line with the literature.

Although we recruited subjects with liver diseases and BMI up to 39, subjects only had liver fat up to about 18% PDFF. There is a known contribution of liver fat to the T1 measurement [29] that is strongly dependent on the readout parameters. Good inter-scanner reproducibility was demonstrated in this population with these parameters with no trend of worse reproducibility with increasing fat fraction, but it is possible that still higher liver fats would show worse reproducibility. Acquisition of MOLLI data with a fat suppression technique is only available on one scanner platform; therefore, similar data could not be taken to measure reproducibility across platform. Other limitations in this study include biases from more 1.5T than 3T phantom and in-vivo reproducibility data, the choice of reference scanner, and limited Philips data. Finally, the MOLLI based technique [17] for T1 mapping used here only sampled 1 slice in each breath-hold. This is a limitation of the readout method, rather than of the technique.

Due to practical limitations, only a small number of participants were evaluated using the Philips scanners at 1.5T and 3T. Although a more balanced sample size per scanner would have been preferable, multiple phantom measurements performed across these scanners showed excellent reproducibility. The ability to standardise across different scanner field strength, manufacturers, and models, is important in the clinical trial setting where accurate and consistent evaluation of key outcomes across treatment interventions and patient groups can be aided by the ability to compare data gathered from multiple sites.

Conclusions

Multiparametric MR-derived metrics, cT1, $T2^*$ and PDFF, have good repeatability and reproducibility that can quantify liver tissue characteristics independent of scanner manufacturer (Philips or Siemens) and field strength (1.5T or 3T). Multiparametric MRI is a non-invasive method that does not require additional hardware, and can be completed in less than 15-minutes, which will have important implications for routine monitoring and assessment of the liver in clinical practice. The ability to standardize metrics will be important in the clinical trial settings for evaluating treatment interventions.

Supporting information

S1 File. MRI scanning sequences.
(DOCX)

S2 File. T1-mapping Functions and precision.
(DOCX)

Acknowledgments

S.N. acknowledges support from the Oxford NIHR Biomedical Research Centre.

Author Contributions

Conceptualization: Jaco Jacobs, Rajarshi Banerjee.

Data curation: Velicia Bachtiar, Henry R. Wilman, Catherine J. Kelly, Katherine E. Groves, Amy H. Herlihy, Carolina C. Fernandes, Mark Halberstadt, Marion Maguire, Naomi Jayaratne, Sophia Linden.

Formal analysis: Henry R. Wilman, Carolina C. Fernandes, Marion Maguire, Naomi Jayaratne, Sophia Linden.

Investigation: Velicia Bachtiar, Rexford Newbould, Katherine E. Groves, Andy McKay, Amy H. Herlihy, Carolina C. Fernandes, Mark Halberstadt.

Methodology: Velicia Bachtiar, Henry R. Wilman, Jaco Jacobs, Catherine J. Kelly, Mark Halberstadt.

Project administration: Velicia Bachtiar, Katherine E. Groves, Andy McKay.

Resources: Jaco Jacobs.

Software: Michael L. Gyngell, Carolina C. Fernandes.

Supervision: Velicia Bachtiar, Matthew D. Kelly, Jaco Jacobs, Rexford Newbould, Catherine J. Kelly, Stefan Neubauer, Rajarshi Banerjee.

Validation: Velicia Bachtiar, Henry R. Wilman, Rexford Newbould.

Visualization: Henry R. Wilman.

Writing – original draft: Velicia Bachtiar.

Writing – review & editing: Velicia Bachtiar, Matthew D. Kelly, Michael L. Gyngell.

References

1. Younossi Z. M., Koenig A. B., Abdelatif D., Fazel Y., Henry L., and Wymer M., "Global epidemiology of nonalcoholic fatty liver disease—Meta-analytic assessment of prevalence, incidence, and outcomes," *Hepatology*, vol. 64, no. 1, pp. 73–84, 2016. https://doi.org/10.1002/hep.28431 PMID: 26707365
2. Loomba R. and Sanyal A. J., "The global NAFLD epidemic," *Nat. Rev. Gastroenterol. Hepatol.*, vol. 10, no. 11, 2013.
3. Poynard T., Ingiliz P., Elkrief L., Munteanu M., Lebray P., Morra R., Messous D., Bismut F. I., Roulot D., Benhamou Y., Thabut D., and Ratziu V., "Concordance in a world without a gold standard: A new non-invasive methodology for improving accuracy of fibrosis markers," *PLoS One*, vol. 3, no. 12, pp. 1–8, 2008.
4. Bedossa P., Dargere D., and Paradis V., "Sampling variability of liver fibrosis in chronic hepatitis C," *Hepatology*, vol. 38, no. 6, p. ajhep09022, 2003.
5. Tapper E. B. and Lok A. S.-F., "Use of Liver Imaging and Biopsy in Clinical Practice," *N. Engl. J. Med.*, vol. 377, no. 8, pp. 756–768, 2017. https://doi.org/10.1056/NEJMra1610570 PMID: 28834467
6. Friedman L. S., "Controversies in liver biopsy: Who, where, when, how, why?," *Curr. Gastroenterol. Rep.*, vol. 6, no. 1, pp. 30–36, Jan. 2004. PMID: 14720451
7. Pavlides M., Banerjee R., Sellwood J., Kelly C. J., Robson M. D., Booth J. C., Collier J., Neubauer S., and Barnes E., "Multiparametric magnetic resonance imaging predicts clinical outcomes in patients with chronic liver disease," *J. Hepatol.*, vol. 64, no. 2, pp. 308–315, 2016. https://doi.org/10.1016/j.jhep.2015.10.009 PMID: 26471505
8. Banerjee R., Pavlides M., Tunnicliffe E. M., Piechnik S. K., Sarania N., Philips R., Collier J. D., Booth J. C., Schneider J. E., Wang L. M., Delaney D. W., Fleming K. A., Robson M. D., Barnes E., and Neubauer S., "Multiparametric magnetic resonance for the non-invasive diagnosis of liver disease," *J. Hepatol.*, vol. 60, no. 1, pp. 69–77, 2014. https://doi.org/10.1016/j.jhep.2013.09.002 PMID: 24036007
9. Pavlides M., Banerjee R., Tunnicliffe E. M., Kelly C., Collier J., Wang L. M., Fleming K. A., Cobbold J. F., Robson M. D., Neubauer S., and Barnes E., "Multiparametric magnetic resonance imaging for the assessment of non-alcoholic fatty liver disease severity," *Liver Int.*, vol. 37, no. 7, pp. 1065–1073, 2017. https://doi.org/10.1111/liv.13284 PMID: 27778429
10. Tunnicliffe E. M., Banerjee R., Pavlides M., Neubauer S., and Robson M. D., "A model for hepatic fibrosis: the competing effects of cell loss and iron on shortened modified Look-Locker inversion recovery T1 (shMOLLI-T1) in the liver," *J. Magn. Reson. Imaging*, vol. 45, no. 2, pp. 450–462, 2017. https://doi.org/10.1002/jmri.25392 PMID: 27448630
11. Blake L., V Duarte R., and Cummins C., "Decision analytic model of the diagnostic pathways for patients with suspected non-alcoholic fatty liver disease using non-invasive transient elastography and

multiparametric magnetic resonance imaging," *BMJ Open*, vol. 6, no. 9, p. e010507, 2016. https://doi.org/10.1136/bmjopen-2015-010507 PMID: 27650757

12. Eddowes P. J., McDonald N., Davies N., Semple S. I. K., Kendall T. J., Hodson J., Newsome P. N., Flintham R. B., Wesolowski R., Blake L., Duarte R. V., Kelly C. J., Herlihy A. H., Kelly M. D., Olliff S. P., Hübscher S. G., Fallowfield J. A., and Hirschfield G. M., "Utility and cost evaluation of multiparametric magnetic resonance imaging for the assessment of non-alcoholic fatty liver disease," *Aliment. Pharmacol. Ther.*, vol. 47, no. 5, pp. 631–644, 2018. https://doi.org/10.1111/apt.14469 PMID: 29271504
13. McDonald N., Eddowes P. J., Hodson J., Semple S. I. K., Davies N. P., Kelly C. J., Kin S., Phillips M., Herlihy A. H., Kendall T. J., Brown R. M., Neil D. A. H., Hübscher S. G., Hirschfield G. M., and Fallowfield J. A., "Multiparametric magnetic resonance imaging for quantitation of liver disease: A two-centre cross-sectional observational study," *Sci. Rep.*, vol. 8, no. 1, pp. 1–10, 2018. https://doi.org/10.1038/s41598-017-17765-5
14. Koo T. K. and Li M. Y., "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, 2016. https://doi.org/10.1016/j.jcm.2016.02.012 PMID: 27330520
15. Bartlett J. W. and Frost C., "Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables," *Ultrasound Obstet. Gynecol.*, vol. 31, no. 4, pp. 466–475, 2008. https://doi.org/10.1002/uog.5256 PMID: 18306169
16. Hines C. D. G., Yu H., Shimakawa A., McKenzie C. A., Brittain J. H., and Reeder S. B., "T1 independent, T2* corrected MRI with accurate spectral modeling for quantification of fat: Validation in a fat-water-SPIO phantom," *J. Magn. Reson. Imaging*, vol. 30, no. 5, pp. 1215–1222, 2009. https://doi.org/10.1002/jmri.21957 PMID: 19856457
17. Messroghli D. R., Radjenovic A., Kozerke S., Higgins D. M., Sivananthan M. U., and Ridgway J. P., "Modified look-locker inversion recovery (MOLLI) for high-resolution T 1 mapping of the heart," *Magn. Reson. Med.*, vol. 52, no. 1, pp. 141–146, 2004. https://doi.org/10.1002/mrm.20110 PMID: 15236377
18. Wilman H. R., Kelly M., Garratt S., Matthews P. M., Milanesi M., Herlihy A., Gyngell M., Neubauer S., Bell J. D., Banerjee R., and Thomas E. L., "Characterisation of liver fat in the UK Biobank cohort," *PLoS One*, vol. 12, no. 2, pp. 1–14, 2017.
19. R. C. Team and others, "R foundation for statistical computing," Vienna, Austria, vol. 3, no. 0, 2013.
20. Pinheiro J., Bates D., DebRoy S., Sarkar D., and R Core Team, "{nlme}: Linear and Nonlinear Mixed Effects Models." 2017.
21. Bernal-Rusiel J. L., Greve D. N., Reuter M., Fischl B., and Sabuncu M. R., "Statistical analysis of longitudinal neuroimage data with Linear Mixed Effects models.," *Neuroimage*, vol. 66, pp. 249–260, Feb. 2013. https://doi.org/10.1016/j.neuroimage.2012.10.065 PMID: 23123680
22. Bartlett J. W., De Stavola B. L., and Frost C., "Linear mixed models for replication data to efficiently allow for covariate measurement error.," *Stat. Med.*, vol. 28, no. 25, pp. 3158–3178, Nov. 2009. https://doi.org/10.1002/sim.3713 PMID: 19777493
23. Harrison S. A., Dennis A., Fiore M. M., Kelly M. D., Kelly C. J., Paredes A. H., Whitehead J. M., Neubauer S., Traber P. G., and Banerjee R., "Utility and variability of three non-invasive liver fibrosis imaging modalities to evaluate efficacy of GR-MD-02 in subjects with NASH and bridging fibrosis during a phase-2 randomized clinical trial," *PLoS One*, vol. 13, no. 9, pp. 8–15, 2018.
24. Trout A. T. and Mahley A. D., "with MR Elastography: Agreement and Repeatability across Imaging," vol. 000, no. 0, pp. 1–12, 2016.
25. Hines C. D. G., Bley T. A., Lindstrom M. J., and Reeder S. B., "Repeatability of magnetic resonance elastography for quantification of hepatic stiffness," *J. Magn. Reson. Imaging*, vol. 31, no. 3, pp. 725–731, 2010. https://doi.org/10.1002/jmri.22066 PMID: 20187219
26. Bane O., Hectors S. J., Wagner M., Arlinghaus L. L., Aryal M. P., Cao Y., Chenevert T. L., Fennessy F., Huang W., Hylton N. M., Kalpathy-Cramer J., Keenan K. E., Malyarenko D. I., Mulkern R. V., Newitt D. C., Russek S. E., Stupic K. F., Tudorica A., Wilmes L. J., Yankeelov T. E., Yen Y. F., Boss M. A., and Taouli B., "Accuracy, repeatability, and interplatform reproducibility of T1quantification methods used for DCE-MRI: Results from a multicenter phantom study," *Magn. Reson. Med.*, vol. 79, no. 5, pp. 2564–2575, 2018. https://doi.org/10.1002/mrm.26903 PMID: 28913930
27. Reeder S. B., Hu H. H., and Sirlin C. B., "Proton Density Fat-Fraction: A Standardized MR-Based Biomarker of Tissue Fat Concentration," *J. Magn. Reson. Imaging*, vol. 36, no. 5, pp. 1011–1014, Nov. 2012. https://doi.org/10.1002/jmri.23741 PMID: 22777847
28. Yokoo T., Serai S. D., Pirasteh A., Bashir M. R., Hamilton G., Hernando D., Hu H. H., Hetterich H., Kühn J.-P., Kukuk G. M., Loomba R., Middleton M. S., Obuchowski N. A., Song J. S., Tang A., Wu X., Reeder S. B., and Sirlin C. B., "Linearity, Bias, and Precision of Hepatic Proton Density Fat Fraction Measurements by Using MR Imaging: A Meta-Analysis," *Radiology*, vol. 286, no. 2, pp. 486–498, Sep. 2017. https://doi.org/10.1148/radiol.2017170550 PMID: 28892458

29. Mozes F. E., Tunnicliffe E. M., Pavlides M., and Robson M. D., "Influence of fat on liver T1 measurements using modified Look–Locker inversion recovery (MOLLI) methods at 3T," *J. Magn. Reson. Imaging*, vol. 44, no. 1, pp. 105–111, 2016. https://doi.org/10.1002/jmri.25146 PMID: 26762615