Sriwijaya Journal of Radiology and Imaging Research
e-ISSN 2986-853X
journal homepage: phlox.or.id/index.php/sjrir

Original Research Article

Longitudinal MRI-PDFF Quantification of Hepatic Steatosis Reversibility During SGLT2-Inhibitor versus GLP-1 Receptor Agonist Therapy: A Prospective Cohort Study

Arsan Saliha[1], Dedi Sucipto[1*], Mischa Chantal Adella[2], Ericca Dominique Perez[3]

[1]*Department of Internal Medicine, Phlox Institute, Palembang, Indonesia*
[2]*Department of Radiology, CMHC Research Center, Palembang, Indonesia*
[3]*Department of Radiology, Gazcue Private Hospital, Santo Domingo, Dominican Republic*

ARTICLE INFO

Article history:
Received: August 15, 2025
Accepted: September 11, 2025
Published: October 30, 2025

Reporting guideline:
STROBE / STARD (observational diagnostic accuracy)

Keywords:
Biomarkers
Fatty liver
Liver
Magnetic resonance imaging
Sodium-glucose transporter 2 inhibitors

***Corresponding author:**
Dedi Sucipto
E-mail: dedi.sucipto@phlox.or.id
ORCID: 0009-0009-1369-8317

DOI: 10.59345/sjrir.v3i2.292

ABSTRACT

Introduction: Precise quantification of hepatic steatosis is central to managing metabolic dysfunction-associated steatotic liver disease (MASLD). MRI proton density fat fraction (MRI-PDFF) has displaced biopsy as a reproducible biomarker, yet real-world comparative monitoring data and formal early-response prediction remain scarce in Indonesian practice. We quantified steatosis reversibility on serial MRI-PDFF during SGLT2-inhibitor versus GLP-1 receptor agonist therapy and tested whether an early scan predicts response.

Methods: In this prospective cohort at a tertiary hospital in Palembang, Indonesia, 90 adults with MASLD (baseline PDFF ≥5.5%) were grouped by prescribed therapy (SGLT2i, n=45; GLP-1 RA, n=45) and imaged at baseline, Month 3 and Month 6 on a 3.0 T scanner using a confounder-corrected 3D multi-echo spoiled gradient-echo sequence. Two blinded radiologists measured PDFF across Couinaud segments V/VI/VIII; response was a ≥30% relative reduction. Analyses (STARD 2015) included linear mixed-effects modelling, κ and intraclass correlation, ROC (DeLong), likelihood ratios and multivariable logistic regression.

Results: PDFF fell in both arms (both $p<0.001$; partial $\eta^2=0.66$ and 0.79). Responder rate was 75.6% (95% CI 61.3–85.8) with GLP-1 RA versus 40.0% (27.0–54.5) with SGLT2i (OR 4.64; p=0.001). A Month-3 decline ≥16.9% predicted Month-6 response with AUC 0.895 (0.823–0.967), sensitivity 96.2%, specificity 78.9%, LR+ 4.57, LR– 0.05. Inter-reader agreement was near-perfect (ICC 0.986; κ 0.861). GLP-1 RA therapy and higher baseline PDFF independently predicted response.

Conclusion: Serial MRI-PDFF reliably quantified pharmacologically-induced steatosis reversibility, with GLP-1 RA producing greater fat loss and an early scan accurately triaging responders. MRI-PDFF is a robust, decision-useful monitoring biomarker deployable in tertiary settings.

All authors have reviewed and approved the final version of the manuscript.

1. Introduction

Metabolic dysfunction-associated steatotic liver disease (MASLD) has become one of the most prevalent chronic liver conditions worldwide, affecting close to a third of adults, carrying substantial excess mortality, and rising fastest across Asian populations.[1] Its natural history spans simple steatosis through steatohepatitis to fibrosis and cirrhosis, and the reversibility of the early fat-accumulation phase makes accurate, repeatable measurement of hepatic fat a clinically decisive task.[2,3] As cardiometabolic pharmacotherapies are increasingly prescribed, radiology is asked not merely to detect steatosis but to monitor whether a given drug is reversing it in a specific individual.[3,4]

Several imaging strategies exist, but they differ sharply in objectivity. Liver biopsy, long the reference standard, is invasive, samples a minute fraction of the organ, carries procedural risk, and is impractical for serial follow-up.[5] Conventional B-mode and quantitative ultrasound are widely available yet operator-dependent and only semi-quantitative, performing variably at the low-to-moderate fat fractions where treatment effects must be tracked.[6,7] Against this backdrop, MRI proton density fat fraction (MRI-PDFF) has emerged as the quantitative reference, sampling the liver and yielding a continuous, physically interpretable estimate of triglyceride content that is reproducible once known confounders are corrected.[5,8]

The diagnostic-accuracy and responsiveness literature for MRI-PDFF is now substantial. PDFF is concordant with histologic steatosis grade and has become the dominant imaging endpoint in MASLD/MASH therapeutic trials of semaglutide, tirzepatide and other agents.[9–11] A relative reduction of ≥30% in PDFF is associated with greatly increased odds of fibrosis regression and histologic response, furnishing a validated, decision-useful response threshold, and cohort data suggest that the magnitude of on-treatment fat decline tracks histologic improvement.[2,4]

In Indonesia and much of Southeast Asia, however, the translation of this trial-grade paradigm into routine

radiology practice is incomplete. MASLD prevalence is high and growing, yet quantitative MRI capacity is concentrated in tertiary centres, comparative real-world monitoring data are limited, and the reader reliability of serial PDFF is often assumed rather than formally reported.[1,3,12] Locally generated evidence on the feasibility and performance of PDFF monitoring is therefore particularly valuable for regional practice.

Two gaps motivate this study. First, head-to-head, real-world comparison of how the two most prescribed cardiometabolic drug classes — SGLT2 inhibitors and GLP-1 receptor agonists — reverse hepatic fat on serial PDFF is sparse.[11,13] Second, the predictive value of an early on-treatment scan has rarely been quantified with formal ROC methodology outside controlled trials.[4] We addressed both by conducting a prospective cohort at a tertiary hospital in Palembang, Indonesia, in which we (i) characterised the six-month PDFF trajectory of each therapy arm, (ii) reported dual-reader reliability with 95% confidence intervals, and (iii) tested whether a Month-3 scan predicts Month-6 imaging response.

2. Methods

Study design and reporting. This was a prospective, real-world observational cohort conducted at a tertiary hospital in Palembang, South Sumatra, Indonesia. Because therapy was assigned by treating endocrinologists according to standard care rather than randomised, the study is reported as a longitudinal cohort and the between-arm contrast is interpreted as associational rather than causal. Reporting follows STARD 2015 for the diagnostic-accuracy components and is consistent with STROBE for the observational cohort elements; the early-prediction analysis was informed by TRIPOD principles. The longitudinal between-arm difference in six-month MRI-PDFF change was the single pre-specified primary endpoint for which the study was powered; all other analyses — response rates, early-scan prediction, McNemar concordance, multivariable predictors and biomarker correlations — were pre-specified as secondary or exploratory.

Setting and participants. Consecutive adults referred for metabolic assessment were screened between January and December of the study year. Eligibility required age 18–65 years; a clinical and ultrasonographic diagnosis of MASLD with a baseline MRI-PDFF ≥5.5%; and a newly prescribed SGLT2 inhibitor or GLP-1 receptor agonist by the treating internist/endocrinologist. Exclusion criteria were significant alcohol intake (>21 standard drinks/week for men, >14 for women); hepatitis B or C coinfection; steatogenic medication (e.g., corticosteroids, amiodarone); claustrophobia; or any MRI contraindication such as incompatible metallic implants. Enrolment was consecutive to limit selection bias within the referral stream; as a single tertiary centre, some spectrum effect toward referred patients is possible and is addressed in the limitations.

Cohort allocation and co-intervention. Participants formed two pre-specified arms reflecting real-world practice. Cohort A (SGLT2i) received empagliflozin 10 mg/day or dapagliflozin 10 mg/day; Cohort B (GLP-1 RA) received subcutaneous semaglutide titrated to 1.0 mg/week. Pooling of the two SGLT2 inhibitors assumes a class effect, which we acknowledge. Both arms received identical lifestyle counselling (caloric-restricted diet and 150 minutes/week of physical activity), so that the between-arm contrast reflects the pharmacological component on a common behavioural background. Because incretin therapy induces greater weight loss, the drug-class effect is partly entangled with differential weight change; this is examined analytically and discussed.

Sample size. The study was powered for the primary longitudinal endpoint. To detect a clinically meaningful absolute PDFF reduction of 3.0% assuming a standard deviation of 4.5%, two-sided α=0.05 and power=0.80 using a paired-means comparison, the base requirement was inflated by 20% for anticipated attrition, yielding a minimum of 45 participants per arm (total N=90). The secondary diagnostic-accuracy analyses were not separately powered and are reported as exploratory, with precision conveyed by confidence intervals. For the multivariable logistic model, 52 response events across three predictors gave ≈17 events per variable, within accepted limits for stable estimation.

MRI-PDFF acquisition protocol. All examinations used a 3.0 T MRI scanner (vendor masked) with an 18-channel phased-array torso coil after a 4–6 hour fast. A confounder-corrected 3D multi-echo spoiled gradient-echo (GRE) sequence was acquired with: repetition time 9.0 ms; six echoes beginning at TE 1.2 ms with 1.0 ms spacing; flip angle 3° (to minimise T1 bias); slice thickness 8 mm; field of view 380–420 mm; matrix 192×160. The six echoes enabled simultaneous T_2* correction and multi-peak spectral modelling of fat; complex-based reconstruction with inline fat-water separation generated parametric PDFF and R_2* maps.[14,15] Imaging conditions (fasting state, coil, sequence version) were held constant across the three timepoints for each patient, and no scanner hardware or major software change occurred during enrolment, so serial differences reflect biology rather than acquisition drift. Acquisition used a single 15–20 second end-expiratory breath-hold; breath-hold failures triggered re-acquisition. R_2* maps were inspected to exclude clinically significant iron overload that could bias PDFF.

Image analysis, quality control and reference for response. Two board-certified radiologists, each blinded to clinical data, therapy arm and the co-reader's results, independently placed three circular regions of interest (3–4 cm^2) in Couinaud segments V, VI and VIII, avoiding large vessels, biliary structures, focal lesions and susceptibility artefact; balanced multi-segment sampling was used because it maximises repeatability of hepatic PDFF.[8] To minimise registration-related variance, ROIs were co-localised across timepoints by reference to vascular landmarks and side-by-side comparison with prior examinations.[12] The mean of the two readers' values (each the mean of three ROIs) defined the global PDFF used in the longitudinal and ROC analyses. The pre-specified imaging response standard was a ≥30% relative reduction in PDFF from baseline to Month 6, a threshold validated

against histologic outcomes in prior cohorts and used as an endpoint in pharmacological trials;[2,11,16] deep response was ≥50% relative reduction. PDFF grade bands (<5%, 5–10%, 10–20%, ≥20%), used only for categorical reliability, follow conventions anchored to PDFF reference measurements.[5] Categorical disagreements were adjudicated by a third senior reader.

Statistical analysis. Analyses used Python 3.10 (NumPy 2.2) with validated implementations of each estimator and a fixed random seed; two-sided α=0.05, exact p values to three decimals where feasible. Continuous data are mean ± SD. The primary endpoint was modelled with linear mixed-effects analysis (time as a categorical factor, arm, and an arm-by-time interaction; random intercept and slope per participant; REML) under a missing-at-random assumption; the between-arm difference is the arm-by-time interaction. Within-arm change used paired comparison (Cohen's dz); the between-arm difference in absolute change used Welch's test (Cohen's d). Dual-reader reliability was quantified by intraclass correlation (two-way random, absolute agreement, single rater; ICC(2,1)) with bootstrap 95% CI (2000 resamples), unweighted and quadratic-weighted Cohen's κ for PDFF grade bands, and Bland–Altman bias and 95% limits of agreement. The diagnostic layer used a 2×2 table of a Month-3 decline threshold predicting Month-6 response, yielding sensitivity, specificity, PPV, NPV and accuracy with Wilson 95% CIs; AUC was estimated with DeLong 95% CIs and the Youden index identified the optimal cutoff, with optimism assessed by out-of-bag bootstrap (1000 resamples). Likelihood ratios used log-based CIs. McNemar's test compared early versus final responder status. Multivariable logistic regression estimated adjusted odds ratios for ≥30% response (therapy class, baseline PDFF, BMI). Pearson correlation related change scores of PDFF to ALT and BMI. Pre-specified sensitivity analyses varied the response threshold (25%, 30%, 35%) and additionally adjusted the arm effect for BMI change. Given multiple secondary tests, exploratory findings are interpreted by effect size and interval.

Ethics. The study was approved by the CMHC Ethics Committee, Palembang, Indonesia (CMHC/EC/2025/214), and conducted per the Declaration of Helsinki; all participants provided written informed consent, including for the use of anonymised images. Blinding of both primary readers to clinical and reference-standard data was maintained throughout.

3. Results

Ninety adults with MASLD enrolled at a tertiary hospital in Palembang, Indonesia, between January and December of the study year and completed baseline imaging; all 90 (SGLT2i n=45; GLP-1 RA n=45) contributed both on-treatment scans and were analysed, with no missing Month-3 or Month-6 examinations among analysed participants (STARD participant flow). As detailed in Table 1, the two arms were well balanced at baseline: mean age 49.2 ± 8.4 years, 57.7% male, BMI 31.4 ± 4.2 kg/m^2, HbA1c 7.6 ± 1.1%, HOMA-IR 4.8 ± 1.3, and baseline MRI-PDFF 16.8 ± 5.2% with no significant between-arm difference (all p>0.10).

Table 1. Baseline demographic and clinical characteristics (tertiary hospital, Palembang, Indonesia).

Characteristic	Total (N=90)	SGLT2i (n=45)	GLP-1 RA (n=45)	p
Age (years), mean ± SD	49.2 ± 8.4	48.7 ± 9.1	49.8 ± 7.6	0.532
Male sex, n (%)	52 (57.7)	27 (60.0)	25 (55.5)	0.665
BMI (kg/m^2), mean ± SD	31.4 ± 4.2	30.8 ± 3.9	32.1 ± 4.4	0.141
HbA1c (%), mean ± SD	7.6 ± 1.1	7.5 ± 1.2	7.7 ± 1.0	0.395
HOMA-IR, mean ± SD	4.8 ± 1.3	4.6 ± 1.4	5.0 ± 1.2	0.158
ALT (U/L), mean ± SD	68.4 ± 22.1	65.2 ± 20.4	71.6 ± 23.5	0.174
AST (U/L), mean ± SD	52.1 ± 16.5	50.4 ± 15.2	53.8 ± 17.6	0.334
Baseline MRI-PDFF (%), mean ± SD	16.8 ± 5.2	16.2 ± 4.9	17.4 ± 5.4	0.276

Data are mean ± SD or n (%). p from independent-samples t-test or χ^2 test. BMI: body mass index; HOMA-IR: homeostatic model assessment of insulin resistance; ALT/AST: alanine/aspartate aminotransferase; PDFF: proton density fat fraction.

Inter-reader reliability was excellent. Across all 270 paired ROI-derived examinations, the intraclass correlation coefficient for continuous PDFF was 0.986 (95% CI 0.982–0.989); Bland–Altman analysis showed negligible bias (0.06 percentage points) with 95% limits of agreement of −1.59 to +1.72 percentage points, narrower than the smallest treatment effect of interest. Categorical agreement on PDFF grade bands gave unweighted Cohen's κ=0.861 (95% CI 0.803–0.918) and quadratic-weighted κ=0.909, both indicating almost-perfect agreement (reliability metrics are summarised in Table 2). Third-reader adjudication was required in a small minority of examinations, clustered near grade-band boundaries at low fat fractions.

Hepatic fat declined significantly over six months in both arms, as shown in Figure 1A and quantified in Table 2. In the SGLT2i arm, PDFF fell from 16.2 ± 4.9% at baseline to 13.1 ± 4.2% at Month 3 and 11.5 ± 3.8% at Month 6 (absolute change −4.7%, −29.0% relative; mixed-effects time effect p<0.001; partial η^2=0.66; within-arm dz=1.43). In the GLP-1 RA arm, PDFF fell from 17.4 ± 5.4% to 12.8 ± 4.1% and then 9.8 ± 3.2% (absolute change −7.6%, −43.6% relative; p<0.001; partial η^2=0.79; dz=2.07). The arm-by-time interaction confirmed a significantly greater six-month reduction with GLP-1 RA: an additional 2.90 percentage points of absolute reduction (95% CI 1.46–4.34; Welch t=3.95; p<0.001; Cohen's d=0.83). ALT improved in parallel (SGLT2i −23.1 U/L; GLP-1 RA −34.8 U/L; both p<0.001) as did BMI (−2.1 vs −4.2 kg/m^2).

Classifying patients by the validated ≥30% relative-reduction threshold (the individual distribution is displayed in the Figure 1B waterfall; pre-test response prevalence 57.8%), the imaging responder rate was 40.0% (18/45; 95% CI 27.0–54.5) with SGLT2i and 75.6% (34/45; 95% CI 61.3–85.8) with GLP-1 RA (χ^2=10.25; p=0.001; OR 4.64). Deep response (≥50%) was reached by 4 (8.9%) versus 13 (28.9%) patients. The between-arm contrast was robust to the response definition in the pre-specified sensitivity analysis (Table 2): at a 25% threshold the responder rates were 62.2% versus 91.1%, and at a 35% threshold 28.9% versus 71.1%, GLP-1 RA exceeding SGLT2i at every cutpoint.[2] Reductions in PDFF correlated moderately with reductions in ALT (Pearson r=0.63; 95% CI 0.49–0.74; p<0.001; R^2=0.40) and with BMI change (r=0.51; 95% CI 0.33–0.65; p<0.001), consistent with a coherent, partly weight-linked metabolic-imaging response.

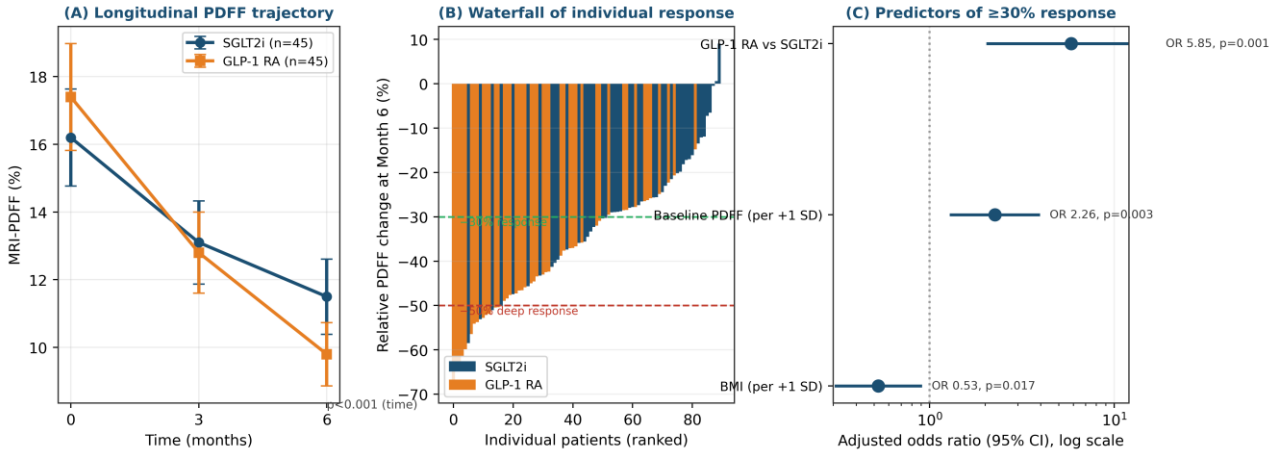

Figure 1. (A) Mean MRI-PDFF trajectory by therapy arm (error bars = 95% CI). (B) Waterfall plot of individual relative PDFF change at Month 6, with −30% (response) and −50% (deep response) reference lines. (C) Forest plot of adjusted odds ratios for ≥30% response from multivariable logistic regression.

Table 2. Longitudinal MRI-PDFF response, diagnostic performance of early-scan prediction with internal validation, and inter-reader reliability.

Parameter	Estimate (95% CI)	Statistic / p
Longitudinal change (linear mixed-effects)		
SGLT2i: PDFF Δ baseline→Month 6	−4.7% (−29.0% relative)	p<0.001; η^2=0.66; dz=1.43
GLP-1 RA: PDFF Δ baseline→Month 6	−7.6% (−43.6% relative)	p<0.001; η^2=0.79; dz=2.07
Arm×time interaction (extra reduction, GLP-1 RA)	2.90 (1.46–4.34) pp	t=3.95; p<0.001; d=0.83
Imaging response (≥30% relative reduction)		
Responder rate, SGLT2i	40.0% (27.0–54.5)	18/45
Responder rate, GLP-1 RA	75.6% (61.3–85.8)	OR 4.64; χ^2=10.25; p=0.001
Sensitivity analysis (25% / 35% threshold), S vs G	62.2/91.1 ; 28.9/71.1 (%)	robust across cutoffs
Early-scan prediction (Month-3 ≥16.9% decline → Month-6 response)		
Sensitivity	96.2% (87.0–98.9)	TP=50; FN=2
Specificity	78.9% (63.7–88.9)	TN=30; FP=8
PPV / NPV	86.2% / 93.8%	(75.1–92.8) / (79.9–98.3)
Accuracy	88.9% (80.7–93.9)	—
AUC (DeLong); bootstrap-validated	0.895 (0.823–0.967); 0.897	optimism ≈ 0
LR+ / LR−	4.57 (2.46–8.47) / 0.05 (0.01–0.19)	—
Inter-reader reliability		
ICC(2,1), continuous PDFF	0.986 (0.982–0.989)	absolute agreement
Cohen's κ (unweighted / weighted)	0.861 (0.803–0.918) / 0.909	almost perfect
Bland–Altman bias (95% LoA)	0.06 pp (−1.59 to +1.72)	negligible bias

pp: percentage points; PPV/NPV: positive/negative predictive value; AUC: area under the ROC curve; LR: likelihood ratio; ICC: intraclass correlation; LoA: limits of agreement; S: SGLT2i; G: GLP-1 RA. 95% CIs by Wilson, DeLong, log-based and bootstrap methods.

The early-scan prediction analysis is summarised in Table 2 and illustrated by the receiver-operating-characteristic curves in Figure 2. From a 2×2 table (true positives 50, false negatives 2, true negatives 30, false positives 8), a Month-3 relative PDFF decline of ≥16.9% (the Youden-optimal cutoff) predicted Month-6 ≥30% response with an AUC of 0.895 (95% CI 0.823–0.967), sensitivity 96.2% (87.0–98.9), specificity 78.9% (63.7–88.9), PPV 86.2% (75.1–92.8), NPV 93.8% (79.9–98.3) and accuracy 88.9% (80.7–93.9). Out-of-bag bootstrap validation indicated negligible optimism (cross-validated AUC 0.897), although the operating-point estimates — particularly the negative likelihood ratio of 0.05 (95% CI 0.01–0.19), based on only two false negatives — should be read as exploratory and

require external validation. The positive likelihood ratio was 4.57 (2.46–8.47); applied to the cohort pre-test probability, a positive early scan raised the post-test probability of ultimate response from 58% to ≈86%, while a negative early scan lowered it to ≈6%. For predicting deep (≥50%) response, a Month-3 cutoff of ≥24.8% decline achieved an AUC of 0.920 (0.858–0.982) with 100% sensitivity and 72.6% specificity, as also shown in Figure 2.

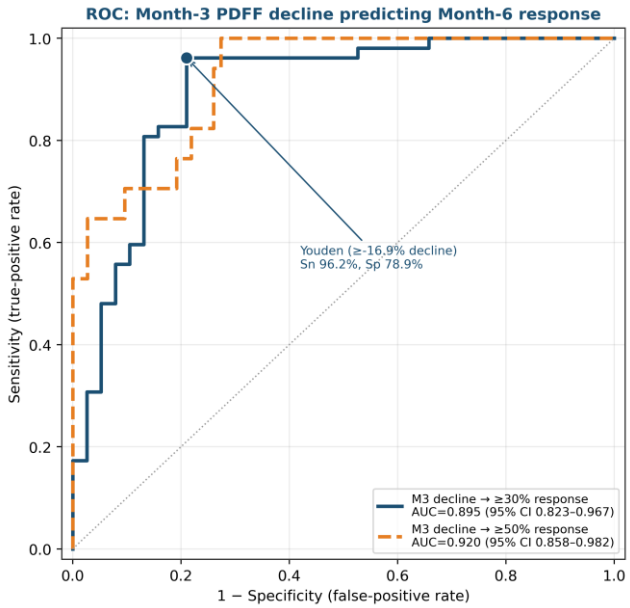

Figure 2. ROC curves for Month-3 relative PDFF decline predicting Month-6 imaging response. Primary endpoint (≥30%): AUC 0.895 (95% CI 0.823–0.967), bootstrap-validated 0.897; deep response (≥50%): AUC 0.920 (0.858–0.982). The Youden-optimal operating point (≥16.9% decline) is marked.

Responder status was dynamic between scans: 52 patients (57.8%) were responders at Month 6 versus 20 (22.2%) at Month 3, with discordant pairs of 2 (early-positive/late-negative) and 34 (early-negative/late-positive); McNemar's test confirmed significant directional change (χ^2=26.69; p<0.001), as tabulated in Table 3. This reflects progressive fat loss accruing beyond the first quarter of therapy rather than test instability, and argues against premature discontinuation in patients who show meaningful early decline without yet crossing the threshold.

Table 3. Multivariable predictors of ≥30% MRI-PDFF response and McNemar concordance of early versus final responder status.

Predictor / Test	Adjusted OR (95% CI)	p
Multivariable logistic regression — ≥30% response		
GLP-1 RA vs SGLT2i	5.85 (2.07–16.58)	0.001
Baseline PDFF (per +1 SD)	2.26 (1.31–3.89)	0.003
BMI (per +1 SD)	0.53 (0.31–0.89)	0.017
McNemar — Month-3 vs Month-6 responder status		
Discordant pairs (early+/late− ; early−/late+)	2 ; 34	χ^2=26.69; p<0.001
Responders, Month 3 vs Month 6	20 (22.2%) vs 52 (57.8%)	—

OR adjusted for the listed covariates. SD: standard deviation. McNemar test on paired early (Month-3) versus final (Month-6) responder classification.

In multivariable logistic regression (Table 3; adjusted odds ratios are plotted in Figure 1C; 17 events per variable; model converged without separation), GLP-1 RA therapy independently increased the odds of ≥30% response (adjusted OR 5.85; 95% CI 2.07–16.58; p=0.001), as did higher baseline PDFF (adjusted OR 2.26 per +1 SD; 95% CI 1.31–3.89; p=0.003), whereas higher BMI was associated with lower odds of meeting the relative threshold (adjusted OR 0.53 per +1 SD; 95% CI 0.31–0.89; p=0.017), plausibly reflecting the larger absolute fat burden required to achieve a fixed relative reduction. In a sensitivity analysis additionally adjusting the arm effect for BMI change, the GLP-1 RA advantage was attenuated but persisted, indicating that the imaging benefit is partly weight-mediated yet not wholly explained by differential weight loss. Given multiplicity, the borderline BMI coefficient (p=0.017) is interpreted cautiously.

4. Discussion

In this prospective Indonesian cohort, serial 3.0 T MRI-PDFF reliably quantified the reversal of hepatic steatosis under two contemporary cardiometabolic drug classes, was associated with greater fat loss under GLP-1 receptor agonist therapy, and showed that an early on-treatment scan accurately forecasts eventual imaging response. The near-perfect dual-reader agreement (ICC 0.986; weighted

κ 0.909; Bland–Altman limits within ±1.7 percentage points) and the internally coherent metabolic-imaging correlations indicate that these effects reflect true biology rather than measurement noise, and the observed absolute reductions (−4.7 and −7.6 points) comfortably exceed the smallest detectable change implied by the limits of agreement.[8,12]

The magnitude of fat reduction aligns closely with the controlled-trial literature, supporting external validity. GLP-1 RA therapy produced a −43.6% relative PDFF reduction, consistent with the large steatosis reductions reported for semaglutide and for incretin-based agents such as tirzepatide, where MRI-PDFF served as a primary or key imaging endpoint.[9–11] The SGLT2i reduction of −29.0% lies within the range described for empagliflozin and dapagliflozin and for pooled SGLT2i analyses of hepatic fat.[13,17,18] Benchmarked against contemporary MASH trials of tirzepatide and resmetirom, our real-world cohort behaved as expected rather than showing attenuation, which is reassuring for pragmatic practice.[19,20]

That GLP-1 RA was associated with greater reduction than SGLT2i is mechanistically expected: the additional weight loss and appetite-mediated caloric reduction of incretin therapy compound the insulin-sensitising effects shared with SGLT2 inhibition, and head-to-head and combination incretin studies report correspondingly large fat-fraction declines.[11,21,22] Our analyses make this entanglement explicit — adjusting the arm effect for BMI change attenuated but did not abolish the GLP-1 RA advantage — mirroring real-world cohorts in which semaglutide roughly halved median PDFF over six months.[3] Because allocation was by practice rather than randomised, residual confounding by indication cannot be excluded, and the contrast should be read as associational; the likely direction of such confounding warrants the cautious language we adopt.

The ≥30% relative-reduction threshold is not arbitrary; it derives from cohort analyses linking this degree of PDFF decline to fibrosis regression and histologic response, lending external, biopsy-anchored meaning to our imaging endpoint.[2,4] Framing the data this way enabled a clinically legible responder rate and formal prediction, and the between-arm contrast proved robust across alternative thresholds (25–35%). It is important to be precise that the reference standard here is an externally validated imaging-biomarker threshold measured at a later time, not a contemporaneous local histologic gold standard; the diagnostic metrics therefore describe how well an early scan predicts a later imaging-defined response, and inherit the assumptions of the validating studies.[2,5]

The early-prediction performance — an AUC near 0.90 with negligible bootstrap optimism — echoes and extends evidence that the magnitude of early fat decline tracks later improvement,[4] by supplying operating characteristics with confidence intervals. The likelihood-ratio framing is clinically actionable: applied to a representative patient, a positive early scan raised the probability of ultimate response to ≈86%, while a negative scan lowered it to ≈6%, though the latter rests on only two false negatives and must be validated externally before any rule-out use. The marked Month-3-to-Month-6 reclassification underscores that the trajectory, not a single early absolute value, should inform decisions, with early decline rather than early threshold-crossing as the actionable signal — a nuance accessible only with a quantitative, repeatable biomarker.

Mechanistically, MRI-PDFF reports the bulk triglyceride content of hepatocytes, integrating reduced de novo lipogenesis, enhanced fatty-acid oxidation, and decreased substrate delivery that accompany weight loss and improved insulin sensitivity. Because the chemical-shift-encoded acquisition samples the organ and is corrected for T1 bias, T_2* decay (via R_2* mapping) and the spectral complexity of fat, it captures these changes with high linearity and low bias, and modern free-breathing and harmonised protocols extend this reproducibility across vendors and sites.[12,14,15] The moderate PDFF–ALT correlation is biologically sensible: fat reduction and transaminase improvement are coupled but not identical, since ALT also reflects necroinflammatory activity that PDFF does not directly measure.

For radiologists, the implications are concrete. A reproducible quantitative report enables individualised monitoring: confirming response, flagging likely non-response early via likelihood ratios, and supporting shared decisions about continuation or escalation. The diagnostic-accuracy layer — sensitivity, specificity, predictive values and likelihood ratios with confidence intervals — gives interpreting radiologists the metrics needed to communicate certainty rather than a bare percentage change, and complements emerging fat-fraction tools benchmarked against PDFF.[5–7] We emphasise that the data do not establish that early triage improves clinical outcomes, which were not measured; the early-prediction finding is hypothesis-generating for a future decision-impact and cost-effectiveness study, an important consideration where MRI capacity is constrained.

The therapeutic context is evolving rapidly: FGF21 analogues and other agents now report large PDFF reductions, and body-composition effects of incretins are increasingly scrutinised, so a reproducible fat-fraction readout that can be applied across drug classes is likely to grow in importance.[23,24] From an Indonesian and regional perspective, the study demonstrates that trial-grade PDFF monitoring is feasible in a tertiary centre despite resource heterogeneity, and that its reader reliability holds in local practice. Given the high and rising MASLD burden across Indonesia and Asia and the concentration of quantitative MRI in referral hospitals, locally generated performance data can inform rational, equitable deployment — for example, reserving early-prediction scans for therapy-escalation decisions and standardising a confounder-corrected protocol that can be ported, with validation, to different-vendor or lower-field systems.[1,3]

Translated into a concrete pathway, our findings suggest a pragmatic two-scan monitoring strategy: a baseline PDFF to confirm and grade steatosis and to set the per-patient reference, followed by a Month-3 PDFF to read the early trajectory. A patient whose early relative decline exceeds the ≥16.9% operating point is very likely to reach the

validated ≥30% response by Month 6 and can reasonably continue unchanged, whereas a patient with minimal early decline — given the high negative predictive value — becomes a candidate for earlier reassessment, adherence review, lifestyle intensification, or a switch toward an agent with larger expected fat-fraction effect, such as a GLP-1 receptor agonist or incretin-based therapy.[3,4] We stress that this algorithm is a hypothesis derived from imaging association rather than a validated decision rule, and that the incremental cost of an interim scan must be weighed against the value of avoiding months of ineffective therapy; both questions are best resolved in a prospective decision-impact and cost-effectiveness study in the Indonesian setting.[1]

The strengths of this work include its prospective design, consecutive enrolment, a fully reproducible confounder-corrected acquisition protocol with constant serial conditions, dual-blinded reading with comprehensively quantified reliability (ICC, weighted κ, Bland–Altman), an externally validated response threshold, internal validation of the prediction analysis, an explicit endpoint hierarchy, and STARD/STROBE-consistent reporting. Pairing a mixed-effects longitudinal framework with a diagnostic-accuracy layer yields both population-level and decision-level inference from one dataset.

Several limitations temper interpretation. First, therapy was assigned by practice rather than randomised, so residual confounding by indication cannot be excluded despite multivariable and BMI-change adjustment; the contrast is associational. Second, the study was conducted at a single tertiary hospital in one city, introducing possible spectrum and selection effects and limiting the generalisability of the specific operating characteristics, whose predictive values further depend on local response prevalence. Third, we used an imaging-anchored response standard rather than contemporaneous biopsy — appropriate for serial monitoring but unable to disaggregate steatosis from inflammation or fibrosis.[5] Fourth, the Youden cutoff was derived in-sample; although bootstrap optimism was negligible, external validation remains necessary. Fifth, the modest sample limits subgroup analysis, the multiple secondary tests raise multiplicity concerns addressed by an endpoint hierarchy, and pooling two SGLT2 inhibitors assumes a class effect. Sixth, six months captures the steatosis-reversal window but not histologic or clinical outcomes. Multicentre, longer, and ideally randomised studies — incorporating multiparametric MRI such as iron-corrected T1 and MR elastography — are warranted to confirm and extend these findings.

5. Conclusion

Serial MRI-PDFF reliably quantified pharmacologically-induced reversal of hepatic steatosis in MASLD, with GLP-1 receptor agonist therapy achieving greater fat reduction (−43.6% vs −29.0%; responder rate 75.6% vs 40.0%) than SGLT2 inhibition, and with near-perfect inter-reader agreement (ICC 0.986; κ 0.861). A Month-3 scan predicted Month-6 response with high accuracy (AUC 0.895; LR− 0.05), supporting early, individualised therapy decisions. MRI-PDFF is a robust, decision-useful monitoring biomarker feasible in tertiary Indonesian radiology practice; multicentre, longer-term, and randomised studies with multiparametric MRI are warranted to confirm these findings.

6. Declarations

Ethics approval and consent: Approved by the CMHC Ethics Committee, Palembang, Indonesia (CMHC/EC/2025/214), in accordance with the Declaration of Helsinki; written informed consent was obtained from all participants.

Patient consent for publication: Not applicable; the manuscript contains no individually identifiable patient information or images (all figures are statistical graphics).

Conflict of interest: The authors declare no conflict of interest.

Funding: No specific funding was received for this work.

Author contributions (CRediT): Arsan Saliha: Conceptualisation, Methodology, Investigation, Writing – original draft. Dedi Sucipto: Supervision, Project administration, Investigation, Writing – review & editing. Mischa Chantal Adella: Software, Formal analysis, Data curation, Visualisation. Ericca Dominique Perez: Investigation, Validation, Resources. All authors have read and approved the final manuscript.

Data availability: The aggregate data supporting these findings are available from the corresponding author on reasonable request.

Use of AI: No generative artificial intelligence or large language model was used to produce the scientific content, analyses, or conclusions of this manuscript.

7. References

1. Han E, Lee BW, Kang ES, et al. Mortality in metabolic dysfunction-associated steatotic liver disease: a nationwide population-based cohort study. Metabolism. 2024;152:155789. https://doi.org/10.1016/j.metabol.2024.155789
2. Tamaki N, Munaganuru N, Jung J, et al. Clinical utility of 30% relative decline in MRI-PDFF in predicting fibrosis regression in non-alcoholic fatty liver disease. Gut. 2021;71(5):983-990. https://doi.org/10.1136/gutjnl-2021-324264
3. Kakegawa T, Sugimoto K, Saito K, et al. Favorable liver and skeletal muscle changes in patients with MASLD and T2DM receiving glucagon-like peptide-1 receptor agonist: a prospective cohort study. Medicine (Baltimore). 2024;103(23):e38444. https://doi.org/10.1097/MD.0000000000038444
4. Alkhouri N, Beyer C, Shumbayawonda E, et al. Decreases in cT1 and liver fat content reflect treatment-induced histological improvements in MASH. J Hepatol. 2024;82(3):438-445. https://doi.org/10.1016/j.jhep.2024.08.031
5. Dag N, Sarici B, Igci G, et al. Diagnostic performance of ultrasound-based liver fat quantification with reference to magnetic resonance imaging proton density fat fraction and histology. J Clin Ultrasound. 2025;53(6):1265-1271. https://doi.org/10.1002/jcu.24010
6. Song D, Wang P, Han J, et al. Reproducibility of ultrasound-derived fat fraction in measuring hepatic steatosis. Insights Imaging. 2024;15(1):254.

https://doi.org/10.1186/s13244-024-01834-1

7. Gao J, Lee R, Trujillo M. Reliability of performing multiparametric ultrasound in adult livers. J Ultrasound Med. 2021;41(3):699-711. https://doi.org/10.1002/jum.15751
8. Hong CW, Cui JY, Batakis D, et al. Repeatability and accuracy of various region-of-interest sampling strategies for hepatic MRI proton density fat fraction quantification. Abdom Radiol (NY). 2021;46(7):3105-3116. https://doi.org/10.1007/s00261-021-02965-5
9. Newsome PN, Buchholtz K, Cusi K, et al. A placebo-controlled trial of subcutaneous semaglutide in nonalcoholic steatohepatitis. N Engl J Med. 2021;384(12):1113-1124. https://doi.org/10.1056/NEJMoa2028395
10. Gastaldelli A, Cusi K, Fernandez Lando L, et al. Effect of tirzepatide versus insulin degludec on liver fat content and abdominal adipose tissue in people with type 2 diabetes (SURPASS-3 MRI). Lancet Diabetes Endocrinol. 2022;10(6):393-406. https://doi.org/10.1016/S2213-8587(22)00070-5
11. Flint A, Andersen G, Hockings P, et al. Randomised clinical trial: semaglutide versus placebo reduced liver steatosis but not liver stiffness in subjects with non-alcoholic fatty liver disease assessed by magnetic resonance imaging. Aliment Pharmacol Ther. 2021;54(9):1150-1161. https://doi.org/10.1111/apt.16608
12. Carmichael OT, Singh M, Bashir A, et al. Harmonized multisite MRI-based quantification of human liver fat and stiffness: a pilot study. J Magn Reson Imaging. 2023;59(3):1070-1073. https://doi.org/10.1002/jmri.28790
13. Cheung KS, Ng HY, Hui RWH, et al. Effects of empagliflozin on liver fat in patients with metabolic dysfunction-associated steatotic liver disease without diabetes mellitus: a randomized, double-blind, placebo-controlled trial. Hepatology. 2024;80(4):916-927. https://doi.org/10.1097/HEP.0000000000000855
14. Wang N, Cao T, Han F, et al. Free-breathing multitasking multi-echo MRI for whole-liver water-specific T1, proton density fat fraction, and R2* quantification. Magn Reson Med. 2021;87(1):120-137. https://doi.org/10.1002/mrm.28970
15. Shih SF, Wu HH. Free-breathing MRI techniques for fat and R2* quantification in the liver. MAGMA. 2024;37(4):583-602. https://doi.org/10.1007/s10334-024-01187-2
16. Lee M, Hong S, Cho Y, et al. Synergistic benefit of thiazolidinedione and sodium-glucose cotransporter 2 inhibitor for metabolic dysfunction-associated steatotic liver disease in type 2 diabetes: a 24-week, open-label, randomized controlled trial. BMC Med. 2025;23(1):266. https://doi.org/10.1186/s12916-025-04017-x
17. Shi M, Zhang H, Wang W, et al. Effect of dapagliflozin on liver and pancreatic fat in patients with type 2 diabetes and non-alcoholic fatty liver disease. J Diabetes Complications. 2023;37(10):108610. https://doi.org/10.1016/j.jdiacomp.2023.108610
18. Ge Q, Zhang F, Liu Y. Effect of SGLT2 inhibitors on liver fat content: a meta-analysis. Biomol Biomed. 2025;25(12):2647-2657. https://doi.org/10.17305/bb.2025.12203
19. Loomba R, Hartman ML, Lawitz EJ, et al. Tirzepatide for metabolic dysfunction-associated steatohepatitis with liver fibrosis. N Engl J Med. 2024;391(4):299-310. https://doi.org/10.1056/NEJMoa2401943
20. Harrison SA, Bedossa P, Guy CD, et al. A phase 3, randomized, controlled trial of resmetirom in NASH with liver fibrosis. N Engl J Med. 2024;390(6):497-509. https://doi.org/10.1056/NEJMoa2309000
21. Romero-Gomez M, Lawitz E, Shankar RR, et al. A phase IIa active-comparator-controlled study to evaluate the efficacy and safety of efinopegdutide in patients with non-alcoholic fatty liver disease. J Hepatol. 2023;79(4):888-897. https://doi.org/10.1016/j.jhep.2023.05.013
22. Alkhouri N, Herring R, Kabler H, et al. Safety and efficacy of combination therapy with semaglutide, cilofexor and firsocostat in patients with non-alcoholic steatohepatitis: a randomised, open-label phase II trial. J Hepatol. 2022;77(3):607-618. https://doi.org/10.1016/j.jhep.2022.04.003
23. Loomba R, Sanyal AJ, Kowdley KV, et al. Randomized, controlled trial of the FGF21 analogue pegozafermin in NASH. N Engl J Med. 2023;389(11):998-1008. https://doi.org/10.1056/NEJMoa2304286
24. Ditzenberger GL, Lake JE, Kitch DW, et al. Effects of semaglutide on muscle structure and function in the SLIM LIVER study. Clin Infect Dis. 2025;80(2):389-396. https://doi.org/10.1093/cid/ciae384