Measurement Is Not the Grade: A Two-Center Whole-Slide Study of AI-Based Lipid-Droplet Area Quantification and NASH CRN Steatosis Grading[a]

Shiau Ru Yang[1], Tsung-Hsuan Wu[1], Hung-Wen Tsai[2], Hung-Chih Chiu[3], Che-Wei Hsu[2], Wei-Lun Alterovitz[4], Pau-Choo Chung[1†], Gil Alterovitz[5†]

[1]*Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan.*

[2]*Department of Pathology, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan.*

[3]*Department of Internal Medicine, National Cheng Kung University Hospital, College of Medicine, National Cheng Kung University, Tainan, Taiwan.*

[4]*Independent Scholar, North Potomac, Maryland, USA.*

[5]*Biomedical Cybernetics Laboratory, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA.*

In digital pathology, artificial intelligence (AI) can turn visually assessed features into continuous measurements. Whether a validated measurement can support a reliable clinical grade is a separate question. Using liver steatosis, we examined whether an AI-derived lipid-droplet area fraction (AI-LDAF) that agreed closely with an expert reference at two centers could be reliably mapped to routine historical Nonalcoholic Steatohepatitis Clinical Research Network (NASH CRN) grades. A segmentation model developed using Center A data was locked before validation and tested in 100 liver-biopsy whole-slide images (50 per center). AI-LDAF and the study-specific expert-derived reference lipid-droplet area fraction (DR-LDAF) were generated through independent pipelines that shared neither lipid-droplet numerators nor evaluable-tissue denominators. Mean per-case intersection over union against a three-expert unanimity-intersection reference mask was 95.73%. AI-LDAF also agreed closely with DR-LDAF (concordance correlation coefficient, 0.9968; root mean square error, 0.65 and 0.74 percentage points at Centers A and B, respectively). Yet among 36 cases historically assigned grade 3 (>66% hepatocyte involvement), neither area fraction exceeded 66%, showing why the two percentage scales should not be treated as equivalent. A scalar AI-LDAF-to-historical-grade mapping calibrated in Center A achieved 76.6% optimism-corrected exact agreement (quadratic-weighted κ, 0.876), but only 48.0% when applied without refitting to Center B (κ, 0.612). The same pattern was observed with DR-LDAF (77.6% versus 50.0%; κ, 0.881 versus 0.617), suggesting that AI-specific segmentation error alone did not readily explain the decline. AI-LDAF remained in close agreement with DR-LDAF at both centers, whereas its mapping to historical grades did not. Native measurement and ordinal projection therefore behaved as separate validation objects across centers: close agreement at the former did not establish transportability of the latter.

Keywords: digital pathology; hepatic steatosis; quantitative AI; NASH CRN grading; clinical validation; cross-center transportability; validation-object mismatch; projection validity.

[†]Email: pcchung@ee.ncku.edu.tw, gil_alterovitz@hms.harvard.edu. [a]This work was supported by the National Science and Technology Council, Taiwan, under Grant NSTC 114-2634-F-006-002.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

1. Introduction

Quantitative medical AI systems are increasingly converting visually assessed or ordinal clinical features into continuous numerical measurements. In digital pathology and related image-based clinical AI, such measurements may improve reproducibility, capture subtle histologic or morphologic variation, and support more precise measurement endpoints. Many established grading standards, however, were developed for human visual assessment. Their thresholds and categories cannot be assumed to translate directly into computational measurement rules. Using such a standard as the reference or target for an AI-native continuous output therefore raises two distinct validation problems. First, the legacy grading construct may assess a different object from that quantified by the native continuous output.[1,2] Second, even when projection of the continuous output onto established ordinal categories is clinically useful, the reader-assigned ordinal target remains subject to intra-reader and inter-reader variability.[3,4] Treating a single observed grade as an error-free ground truth can therefore conflate measurement error, projection error, and reader variability.

Here, liver steatosis provides a concrete example of this problem. Liver-biopsy assessment remains important in metabolic dysfunction-associated steatotic liver disease (MASLD) and metabolic dysfunction-associated steatohepatitis (MASH), particularly in research and therapeutic-development settings. Histologic steatosis has traditionally been communicated through semiquantitative ordinal categories, including the NASH Clinical Research Network (NASH CRN) thresholds of <5%, 5%–33%, 34%–66%, and >66%.[5,6] These thresholds refer to the visually estimated proportion of hepatocytes affected by macrovesicular steatosis, whereas an area-based system measures lipid-droplet occupancy relative to evaluable tissue. These categories are clinically familiar and useful for communication, but they compress continuous histologic variation into broad bins and were not designed as native outputs of AI-based continuous measurement systems. Their numerical boundaries are embedded in a human visual grading procedure, not defined as stand-alone machine thresholds. Applying them directly to an AI-native measurement therefore assumes that the two percentage scales are interchangeable, an assumption that must be demonstrated rather than inferred from their shared percentage units.

Prior work has shown that liver steatosis can be quantified using computational image analysis, including our previously developed BSMatch system evaluated here, and that AI-assisted tools can support MASH histologic assessment.[3,7,8] Morphometric studies have also shown that visual estimates and area-based measurements can diverge.[1,2,9] AIM-MASH further illustrates that proportionate-area quantification and CRN-aligned grading are distinct outputs: segmentation models generate quantitative measurements, whereas dedicated grading models generate CRN-aligned scores.[10] These grading outputs have been evaluated against consensus or panel-based reference standards established by expert pathologists.[3,10] The unresolved question is whether a continuous measurement that performs consistently across centers can also be projected reliably onto historical reader-assigned grades across centers. To our knowledge, this has not been directly tested using a mapping calibrated at one center and then applied, without refitting, to an independent cohort from another center.

To address this question, we distinguished three objects that are often conflated: the historical reader-assigned NASH CRN grade, the expert-derived continuous lipid-droplet area fraction (DR-LDAF), and the AI-derived continuous lipid-droplet area fraction (AI-LDAF). We first evaluated whether AI-LDAF agreed with the same-object DR-LDAF reference and whether this measurement

agreement was maintained across centers. We then tested whether a scalar area-to-grade mapping, calibrated in Center A, remained transportable to Center B without refitting. Repeating the mapping analysis with DR-LDAF as a same-object control helped distinguish instability in the area-to-grade projection from AI-specific segmentation error. This design treated pixel-level segmentation, native continuous measurement, and downstream ordinal projection as separate validation objects. It directly tested whether close cross-center measurement agreement could coexist with a locked ordinal mapping that failed to transport.

Although liver steatosis serves as the empirical example, this distinction may also be relevant to other pathology settings in which quantitative image-derived measurements coexist with reader-assigned ordinal systems, including liver fibrosis staging[10] and HER2 immunohistochemical (IHC) scoring.[11,12] Accounting for the relationship between the native quantitative measurement and its categorical projection may help ensure that clinical validation, intended-use claims, risk controls, and regulatory evidence are appropriately aligned with the validated output.

2. Methods

2.1. *Study design and dataset*

This retrospective two-center validation study evaluated a previously developed AI segmentation system at National Cheng Kung University Hospital (NCKUH, Center A) and Taichung Veterans General Hospital (TVGH, Center B) (Fig. 1). Center A provided the within-center validation cohort and was used to calibrate the scalar area-to-historical-grade mapping, whereas Center B provided the independent cross-center cohort to which the model and Center A-derived mapping were applied without refitting. The validation cohorts were independent of and non-overlapping with the model-development dataset. The study was approved by the institutional review boards of National Cheng Kung University Hospital (NCKUH; IRB No. A-ER-111-101) and Taichung Veterans General Hospital (TVGH; TCVGH-IRB No. CE24356A). All data used in the present analysis were retrospectively collected and de-identified before analysis.

The present study used the previously reported BSMatch segmentation algorithm. The published report described its architecture, patch-sampling strategy, and initial technical benchmarking.[8] For the implementation evaluated here, model development used Center A data exclusively, comprising 10 training whole-slide images and 10 internal test whole-slide images, from which 2,400 training patches and 800 test patches of 512 × 512 pixels were sampled. After model development and internal testing, the trained model and inference pipeline were locked before initiation of the present validation study. No image from Center B was used for training, model selection, parameter tuning, or recalibration. The study design was consistent with the test-data-independence principle described in the 2025 FDA draft guidance for AI-enabled device software functions.[13]

2.2. *Study data and eligibility criteria*

The validation cohort comprised de-identified liver-biopsy whole-slide images and corresponding histopathologic data from adults evaluated for suspected metabolic dysfunction-associated steatotic liver disease at Center A between January 1, 2021, and March 1, 2022, and at Center B between January 1, 2021, and December 31, 2022. Eligible patients were aged 20 years or older and had an available formalin-fixed, paraffin-embedded liver-biopsy specimen stained with hematoxylin and

eosin (H&E) and digitized at 40× magnification. Both centers routinely used PANNORAMIC 250 FLASH III scanners (3DHISTECH Ltd.), and slides were scanned independently at each institution using its local clinical digitization and quality-control workflow.

Source records were screened before random sampling. We screened 1,442 records at Center A, excluded 1,236 records with tumor-related diagnoses, including those coded or described as cancer, tumor, or hepatocellular carcinoma (HCC), and retained 206 cases after record-level screening; the corresponding counts at Center B were 641, 429, and 212 (Table 1). From each retained pool, we selected 56 validation candidates by simple random sampling without stratification by historical grade. We then excluded six per center because of unacceptable image quality or aggregate biopsy tissue length below 1.5 cm,[14] leaving 50 cases per center.

The validation cohort size followed the prespecified sample-size calculation in the source clinical-validation plan, which required 48 evaluable cases per center based on 80% power, a one-sided α of 0.025, and a margin of 12.72 percentage points for segmentation mIoU. Applying this target separately to each center yielded 96 evaluable cases overall; 56 cases per center (112 total) were sampled to accommodate anticipated post-sampling exclusions. No separate sample-size calculation was performed for the exploratory area-to-grade projection analysis.

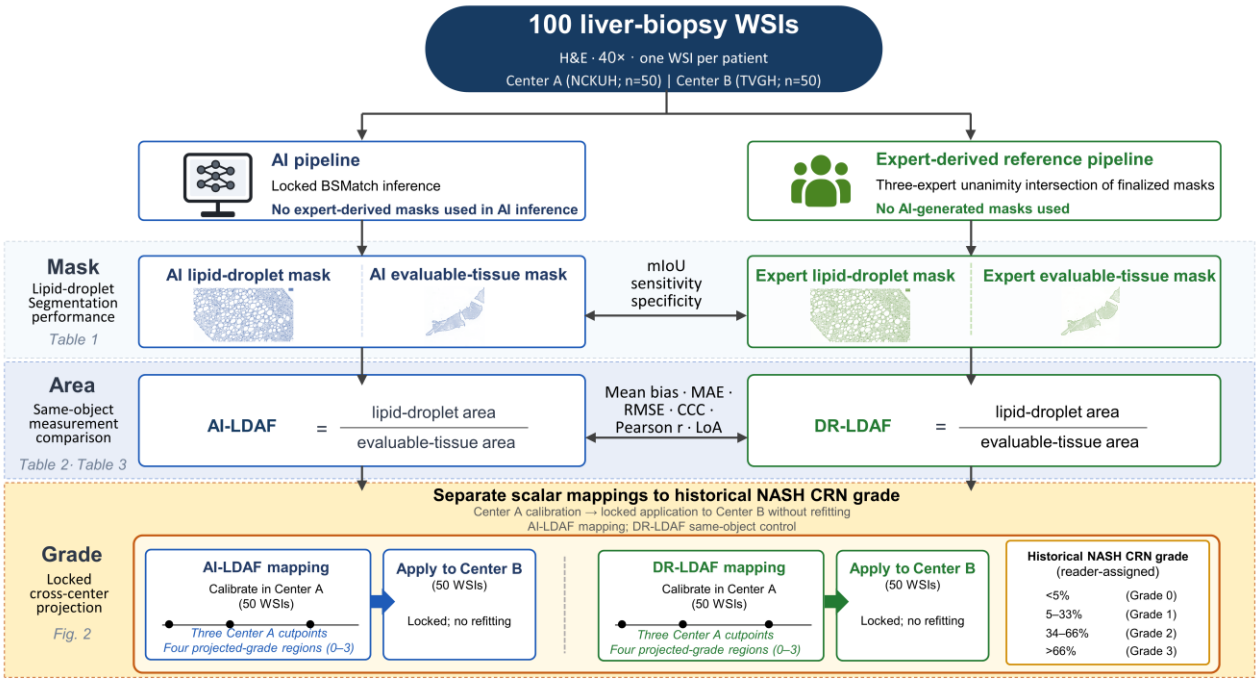


Fig. 1. Study design and separation of validation objects. One 40× H&E whole-slide image from each of 100 patients (50 per center) entered independent AI and expert-reference pipelines with no shared masks. Medical-student annotators manually prepared lipid-droplet pre-annotations without access to BSMatch output; three experts independently finalized separate masks, whose 3/3 pixel-wise intersections defined the reference. Segmentation metrics assessed mask agreement, whereas AI-LDAF and DR-LDAF were evaluated as continuous measurements. Separate Center A mappings to historical grade were locked and applied to Center B, with DR-LDAF as a same-object control.

2.3. *AI system*

The lipid-droplet segmentation model evaluated in this study was the previously developed Boundary Segmentation and Matching (BSMatch) system.[8] BSMatch was designed to reduce false-positive detection of visually similar non-droplet structures by combining region and boundary information. Its two-branch RnB-UNet predicts lipid-droplet region and boundary probability maps, with a boundary-matching loss encouraging consistency between the two outputs. A watershed step separates connected droplets, after which the boundary-matching score filters candidate regions that are poorly matched to their predicted boundaries. The algorithmic architecture, training procedure, loss formulation, and comparative technical evaluation are described in the prior report; the present study focuses on locked-model whole-slide validation and cross-center transportability. Whole-slide inference was performed on 40× images acquired with a 3DHISTECH PANNORAMIC 250 FLASH III scanner. Each whole-slide image was processed as 512 × 512-pixel patches, and patch-level predictions were recombined into whole-slide masks. For every validation whole-slide image (WSI), the AI independently generated a lipid-droplet mask and an evaluable-tissue mask. AI-LDAF was calculated entirely from these AI-generated masks: the numerator was the AI-segmented lipid-droplet area, and the denominator was the AI-segmented evaluable-tissue area after automated exclusion of portal regions, dilated luminal spaces, including dilated central veins and sinusoids, non-tissue background, and prespecified artifacts. No expert-reader annotation was used during AI inference or in either component of AI-LDAF. The software also generated a visual lipid-droplet overlay.

The output evaluated in this study was continuous lipid-droplet area quantification. The AI system did not output an automated NASH CRN grade or independently determine a diagnosis. The AI output was not used for clinical diagnosis or decision-making in this retrospective study. Nominal NASH CRN category boundaries were used only for descriptive post hoc binning. Any future NASH CRN-compatible categorical output would require a separately developed categorical model evaluated against a prespecified reference standard established on the same H&E section used for quantitative analysis, followed by locked external validation.

2.4. *Expert-derived reference masks and DR-LDAF*

For the present validation study, the reference masks were produced in two stages. To make exhaustive whole-slide annotation feasible given the large number of overlapping lipid droplets, many with missing or weak boundaries,[7] medical-student annotators trained in the study protocol first manually prepared a lipid-droplet pre-annotation for each source WSI, without access to BSMatch output or other automated segmentation. Three senior expert readers, each with more than 8 years of relevant liver-disease experience, then independently reviewed each complete WSI together with its pre-annotation and finalized separate reader-specific masks, adding missed lipid-droplet regions, removing false-positive regions, and revising boundaries as needed. The reader panel comprised two pathologists with diagnostic pathology expertise and one senior attending gastroenterologist. All three followed the same prespecified written annotation protocol, literature-informed morphologic criteria for hepatic steatosis,[5,6] and a sequential workflow for portal areas, lipid droplets, dilated central veins and sinusoids, non-tissue background, and prespecified artifacts. Although all three readers began from the same lipid-droplet pre-annotation prepared by the medical-student annotators for each WSI, they had no access to one another’s edits or final masks.

The final reader-specific lipid-droplet and evaluable-tissue masks were intersected separately; a pixel was retained only when annotated by all three readers, constituting a 3/3 unanimity rule. DR-LDAF was calculated as the intersection-derived lipid-droplet area divided by the corresponding intersection-derived evaluable-tissue area after the same prespecified exclusions; no AI-generated mask contributed to either component. Neither the medical-student pre-annotations nor the expert edits were used for model training, fine-tuning, model selection, threshold adjustment, recalibration, or any other modification of the locked BSMatch model. AI-LDAF and DR-LDAF therefore implemented the same continuous measurement definition through independent pipelines that shared neither the lipid-droplet numerator nor the evaluable-tissue denominator. Inter-reader reliability of the three final reader-specific LDAF measurements was assessed using the intraclass correlation coefficient [ICC(2,1)].[15,16] The use of a three-reader expert panel was consistent with prior pathology-AI validation studies using multi-expert reference standards or expert review of model-generated histologic overlays.[3,17]

Historical NASH CRN steatosis grades were pre-existing routine biopsy-level assessments made by one or more readers under each center's routine diagnostic workflow; the number and identity of readers may have varied across cases. Each historical record was linked to the same biopsy specimen as the study WSI, but correspondence to the exact H&E slide or histologic section represented by that WSI could not be verified. The recorded grades were therefore treated as prior ordinal records rather than study-specific consensus labels. The 5%, 33%, and 66% boundaries came from the established NASH CRN visual grading categories, not the present data,[5,6] and were used only as nominal post hoc bins because they reflect semiquantitative hepatocyte involvement rather than morphometric lipid-droplet area.

2.5. *Segmentation and measurement metrics*

Pixel-level lipid-droplet segmentation performance was assessed by directly comparing the AI-generated lipid-droplet mask with the three-expert unanimity-intersection lipid-droplet reference mask across all pixels in each reconstructed WSI. The primary segmentation endpoint was mean intersection over union (mIoU), calculated as the unweighted mean of the lipid-droplet-region IoU and non-droplet-region IoU.[18]

$$\mathrm{mIoU} = \frac{\mathrm{IoU}_{\mathrm{droplet}} + \mathrm{IoU}_{\mathrm{non\text{-}droplet}}}{2} \quad (1)$$

Per-case sensitivity and specificity were calculated against the same reference mask, with reference lipid-droplet pixels treated as positives and all remaining pixels treated as negatives.

Case-level agreement between AI-LDAF and DR-LDAF was assessed using mean bias, defined as AI-LDAF minus DR-LDAF, mean absolute error (MAE), root mean square error (RMSE), Lin's concordance correlation coefficient (CCC), and Bland–Altman 95% limits of agreement (LoA). LoA were calculated as the mean bias ± 1.96 standard deviations of the paired case-level differences.[19-22] Positive bias indicated higher AI-LDAF than DR-LDAF. Bias, MAE, RMSE, and LoA are reported in percentage points; CCC is dimensionless. Pearson's correlation coefficient was also reported to describe linear association, not as sole evidence of numerical agreement.

Per-case mIoU, sensitivity, and specificity were summarized as arithmetic means within each center and overall. Ninety-five percent confidence intervals were estimated using 5,000 case-level

percentile bootstrap resamples; for the combined cohort, cases were resampled independently within each center to preserve the 1:1 center composition. Continuous-agreement statistics, except LoA, used the same bootstrap scheme. Center B minus Center A differences in RMSE, mean bias, and CCC were estimated by independent within-center resampling and were not interpreted as formal evidence of equivalence because no between-center equivalence margin was prespecified.

Pixel-level segmentation, continuous-measurement agreement, and ordinal projection were treated as separate validation objects and evaluated on their native scales. Evidence at one layer was not treated as validation of another. In particular, no ordinal inter-reader variability criterion was translated into an mIoU or area-percentage threshold.

2.6. *Descriptive binning by nominal CRN category boundaries*

AI-LDAF and DR-LDAF were converted to percentages and post hoc assigned to nominal CRN bins to compare category assignments between two same-object continuous measurements:

$$B(x) = \begin{cases} 0, & x < 5\% \\ 1, & 5\% \le x \le 33\% \\ 2, & 33\% < x \le 66\% \\ 3, & x > 66\% \end{cases} \quad (2)$$

For each case, let A_i and R_i denote AI-LDAF and DR-LDAF, respectively, expressed as percentages. Absolute ordinal displacement was defined as:

$$d_i = |B(A_i) - B(R_i)| \quad (3)$$

Exact-bin agreement, within-one-bin agreement, major discordance (≥2 bins), and mean absolute ordinal displacement were reported. This descriptive analysis was not treated as a validated projection to historical reader-assigned NASH CRN grade.

2.7. *Cross-center transportability of locked LDAF-to-historical-grade mappings*

For AI-LDAF, we calibrated the simplest area-to-grade projection, a scalar cutpoint mapping to historical NASH CRN grade, using Center A only. Candidate cutpoints were defined as midpoints between consecutive distinct Center A AI-LDAF values, and all ordered triplets were exhaustively evaluated. For each triplet, AI-LDAF values were projected to grades 0–3, and the triplet maximizing quadratic-weighted kappa with the historical grade was selected.[23] Ties were resolved sequentially by higher exact agreement, lower ordinal mean absolute error, and finally the lexicographically smallest triplet. The selected mapping was then locked and applied to Center B without refitting; no Center B case entered candidate-cutpoint generation, cutpoint selection, or tie resolution.

Because the same Center A sample was used for cutpoint selection and apparent performance estimation, internal optimism and cutpoint stability were assessed using 5,000 case-level bootstrap resamples, with the complete candidate-generation, cutpoint-search, and tie-breaking procedure repeated within each resample.[24,25] For the optimism-corrected performance metrics reported here, namely exact agreement and quadratic-weighted kappa, optimism was defined as the performance of the bootstrap-derived mapping in the resample in which it was calibrated minus its performance when applied to the original Center A cohort; the optimism-corrected estimate was calculated as the apparent estimate minus mean optimism. Within-one-grade agreement and ordinal mean absolute error were reported using apparent Center A estimates. Cutpoint stability was summarized using the

empirical 2.5th and 97.5th percentiles of the bootstrap-selected values for each cutpoint; these percentiles were interpreted as bootstrap stability ranges rather than formal confidence intervals.

Center B point estimates were obtained by applying the mapping calibrated in the original Center A cohort, without modification, to the complete Center B cohort. To propagate both calibration-sample and external-sample uncertainty, 5,000 bootstrap replicates were generated by independently resampling cases with replacement within each center. In each replicate, the complete mapping procedure was recalibrated using only the Center A resample and applied unchanged to the corresponding Center B resample. Exact agreement, within-one-grade agreement, quadratic-weighted kappa, and ordinal mean absolute error were reported. Cases were classified as exact, downward, or upward mappings when the projected grade was equal to, lower than, or higher than the historical grade, respectively. AI-LDAF and DR-LDAF distributions were summarized descriptively by center and historical grade using medians, and the between-center median difference at historical grade 3 was calculated separately for each measure. Percentile-bootstrap 95% confidence intervals were calculated for the Center B performance metrics. Dedicated bootstrap procedures for interval estimation after optimism correction have been described.[26] Because no such procedure was implemented in the present analysis, optimism-corrected Center A estimates for exact agreement and quadratic-weighted kappa, together with the corresponding differences calculated as Center B minus optimism-corrected Center A, were reported as point estimates only.

The analysis was repeated with DR-LDAF as an expert-derived same-object control to assess whether AI-specific segmentation error alone could explain any between-center difference in projection agreement. The mapping was treated as an exploratory transportability stress test rather than as a validated CRN grading output.

Table 1. Case selection, center-stratified cohort characteristics, and segmentation performance. Tumor-related exclusions included records coded or described as cancer, tumor, or hepatocellular carcinoma (HCC). Sampling was not stratified by historical grade. Post-sampling exclusions reflected unacceptable image quality or aggregate biopsy tissue length <1.5 cm. Age was available in prespecified categories. All included cases met the prespecified image-quality and ≥1.5-cm tissue-length criteria. *Three decimal places were retained for specificity estimates that would otherwise round to 100.00%. IQR, interquartile range.

Characteristic	Center A (NCKUH, n = 50)	Center B (TVGH, n = 50)	Combined (n = 100)
Data selection			
Screened/tumor-related records excluded/retained, n	1,442/1,236/206	641/429/212	2,083/1,665/418
Sampled/excluded after review/included, n	56/6/50	56/6/50	112/12/100
Cohort characteristics			
Age 25–44/45–59/≥60 years, n	15/20/15	17/17/16	32/37/31
Male/female, n	29/21	27/23	56/44
Historical grade 0/1/2/3, n	4/16/16/14	4/7/17/22	8/23/33/36
AI-LDAF, %, median (IQR)	12.01 (5.47–19.33)	9.72 (5.36–16.77)	10.37 (5.39–17.85)
DR-LDAF, %, median (IQR)	12.54 (5.93–19.89)	10.32 (5.66–17.65)	10.88 (5.69–18.55)
Segmentation performance			
mIoU, %	95.52 (94.84–96.13)	95.93 (94.74–96.77)	95.73 (95.06–96.25)
Sensitivity, %	97.42 (96.96–97.83)	98.73 (97.52–99.57)	98.08 (97.42–98.55)
Specificity, %	99.71 (99.66–99.75)	99.995 (99.992–99.997)*	99.85 (99.83–99.87)

3. Results

3.1. *Validation cohort*

At each center, 6 of 56 randomly sampled validation candidates were excluded after image-quality and tissue-length review, leaving 50 cases per center (100 whole-slide images in total, one per patient). Center-specific selection and baseline characteristics are summarized in Table 1. The relatively small number of grade 0 cases may reflect the clinically enriched nature of the cohort, because liver biopsy was generally performed in patients with suspected liver disease rather than in an unselected population.

3.2. *Segmentation performance*

Mean per-case mIoU was 95.73% (95% bootstrap CI, 95.06%–96.25%) overall and was numerically similar across centers (Table 1). Because mIoU weights lipid-droplet and non-droplet IoU equally, the larger non-droplet class is not given greater weight. Mean sensitivity and specificity were 98.08% and 99.85% overall, with center-specific values in Table 1. ICC(2,1) point estimates for final reader-specific LDAF measurements indicated good-to-excellent inter-reader reliability: 0.8926 at Center A, 0.8863 at Center B, and 0.9128 overall.

3.3. *Continuous-measurement agreement*

AI-LDAF and DR-LDAF showed close agreement at the whole-slide level. In the combined cohort, CCC was 0.9968 (95% CI, 0.9957–0.9977), mean bias, defined as AI-LDAF minus DR-LDAF, was −0.521 percentage points (95% CI, −0.618 to −0.437), and the LoA were −1.430 to 0.389 percentage points. Center-specific estimates are reported in Table 2. The Center B-minus-Center A differences were 0.083 percentage points for RMSE (95% CI, −0.201 to 0.404), −0.089 percentage points for mean bias (95% CI, −0.274 to 0.081), and −0.0012 for CCC (95% CI, −0.0032 to 0.0007).

Table 2. Whole-slide continuous-measurement agreement between AI-LDAF and DR-LDAF, overall and by center. Values are point estimates with 95% CIs in parentheses, except for the LoA, which are reported as the lower and upper 95% limits. Mean bias was defined as AI-LDAF minus DR-LDAF; negative values indicate lower AI-LDAF. Center differences were calculated as Center B minus Center A. Dashes indicate that the between-center difference was not calculated. pp, percentage points; CI, confidence interval; MAE, mean absolute error; RMSE, root mean square error; LoA, limits of agreement; CCC, concordance correlation coefficient.

Metric	Center A (n = 50)	Center B (n = 50)	Combined (n = 100)	Difference, B − A (95% CI)
Mean bias, pp	−0.476 (−0.604 to −0.356)	−0.565 (−0.712 to −0.451)	−0.521 (−0.618 to −0.437)	−0.089 (−0.274 to 0.081)
MAE, pp	0.499 (0.389 to 0.624)	0.570 (0.457 to 0.719)	0.535 (0.455 to 0.629)	—
RMSE, pp	0.653 (0.476 to 0.825)	0.736 (0.530 to 1.002)	0.696 (0.551 to 0.866)	0.083 (−0.201 to 0.404)
LoA, pp	−1.360 to 0.408	−1.500 to 0.370	−1.430 to 0.389	—
CCC	0.9973 (0.9960 to 0.9984)	0.9961 (0.9945 to 0.9976)	0.9968 (0.9957 to 0.9977)	−0.0012 (−0.0032 to 0.0007)
Pearson's r	0.9992 (0.9988 to 0.9996)	0.9993 (0.9989 to 0.9998)	0.9992 (0.9989 to 0.9996)	—

3.4. *Comparison using nominal CRN category boundaries*

After post hoc application of the nominal NASH CRN category boundaries described in Sections 2.4 and 2.6, exact-bin agreement between AI-LDAF and DR-LDAF was 98.0%, within-one-bin agreement was 100.0%, major discordance was 0.0%, and mean absolute ordinal displacement was 0.02 bin units (Table 3). The two discordant cases were adjacent boundary crossings: AI-LDAF and DR-LDAF straddled the 5% boundary at 4.79% and 5.21%, respectively, and the 33% boundary at 31.21% and 33.39%, corresponding to absolute differences of 0.42 and 2.18 percentage points.

The resulting area-based category distributions differed markedly from the historical reader-assigned grade distribution. Historical records included 36 grade 3 cases, whereas neither LDAF exceeded 66% in any analyzed WSI. In addition, 99% of AI-LDAF values and 98% of DR-LDAF values were at or below 33%. Thus, direct post hoc application of the nominal NASH CRN category boundaries to the continuous area measurements produced a category distribution substantially different from that of the historical grades.

Table 3. Cross-tabulation of AI-LDAF and DR-LDAF after post hoc application of nominal NASH CRN category boundaries. Values are numbers of cases. Nominal bins were defined as <5%, 5%–33%, >33%–66%, and >66%. AI-LDAF and DR-LDAF were assigned to the same bin in 98 of 100 cases; the two discordant cases differed by one adjacent bin.

AI-LDAF bin	DR-LDAF				Total
	<5%	5%–33%	>33%–66%	>66%	
<5%	21	1	0	0	22
5%–33%	0	76	1	0	77
>33%–66%	0	0	1	0	1
>66%	0	0	0	0	0
Total	21	77	2	0	100

3.5. *Cross-center transportability of locked LDAF-to-historical-grade mappings*

Native-measurement performance. Native-measurement point estimates were numerically similar across centers for both segmentation and continuous-measurement agreement (Tables 1 and 2). We therefore evaluated locked external application of scalar LDAF-to-historical-grade mappings.

AI-LDAF mapping. Center A calibration selected cutpoints of 2.88%, 8.79%, and 19.23%. Apparent exact agreement, within-one-grade agreement, ordinal mean absolute error, and quadratic-weighted kappa were 84.0%, 100.0%, 0.16 grade units, and 0.913, respectively; optimism-corrected exact agreement and kappa were 76.6% and 0.876. In Center B, the corresponding values were 48.0% (95% CI, 32.0%–62.0%), 90.0% (68.0%–98.0%), 0.62 grade units (0.42–0.96), and 0.612 (0.320–0.756). Relative to optimism-corrected Center A, exact agreement and kappa differed by −28.6 percentage points and −0.264.

DR-LDAF same-object control. Center A calibration selected cutpoints of 3.00%, 9.19%, and 19.67%. Apparent exact agreement, within-one-grade agreement, ordinal mean absolute error, and kappa were again 84.0%, 100.0%, 0.16 grade units, and 0.913; optimism-corrected exact agreement and kappa were 77.6% and 0.881. In Center B, the corresponding values were 50.0% (95% CI, 32.0%–64.0%), 90.0% (68.0%–98.0%), 0.60 grade units (0.42–0.96), and 0.617 (0.320–0.756).

Relative to optimism-corrected Center A, exact agreement and kappa differed by −27.6 percentage points and −0.264.

Cutpoint stability. Bootstrap stability ranges for the three cutpoints were 1.00%–3.49%, 5.40%–13.19%, and 14.82%–20.50% for AI-LDAF and 1.13%–3.66%, 5.70%–13.66%, and 17.98%–20.88% for DR-LDAF, with the middle cutpoint varying most.

Direction of discordance. Fig. 2 shows case-level values, medians, and mapping directions relative to the locked cutpoints. Each Center A mapping had six downward and two upward discordances. In Center B, all discordances were downward: 26 for AI-LDAF, including 24 among historical grades 2 or 3, and 25 for DR-LDAF, including 23 among grades 2 or 3.

Grade-specific distributions. Baseline historical-grade distributions differed between centers: Center B had fewer grade 1 cases than Center A (7 versus 16) and more grade 3 cases (22 versus 14) (Table 1). Median AI-LDAF at Center B versus Center A was 8.6% versus 13.2% for historical grade 2 and 15.4% versus 25.7% for grade 3; the corresponding DR-LDAF medians were 9.6% versus 13.8% and 16.0% versus 26.5%. Under the fixed Center A cutpoints, these lower Center B distributions produced more downward mappings but did not establish why the distributions differed. The DR-LDAF mapping showed a parallel external decline.

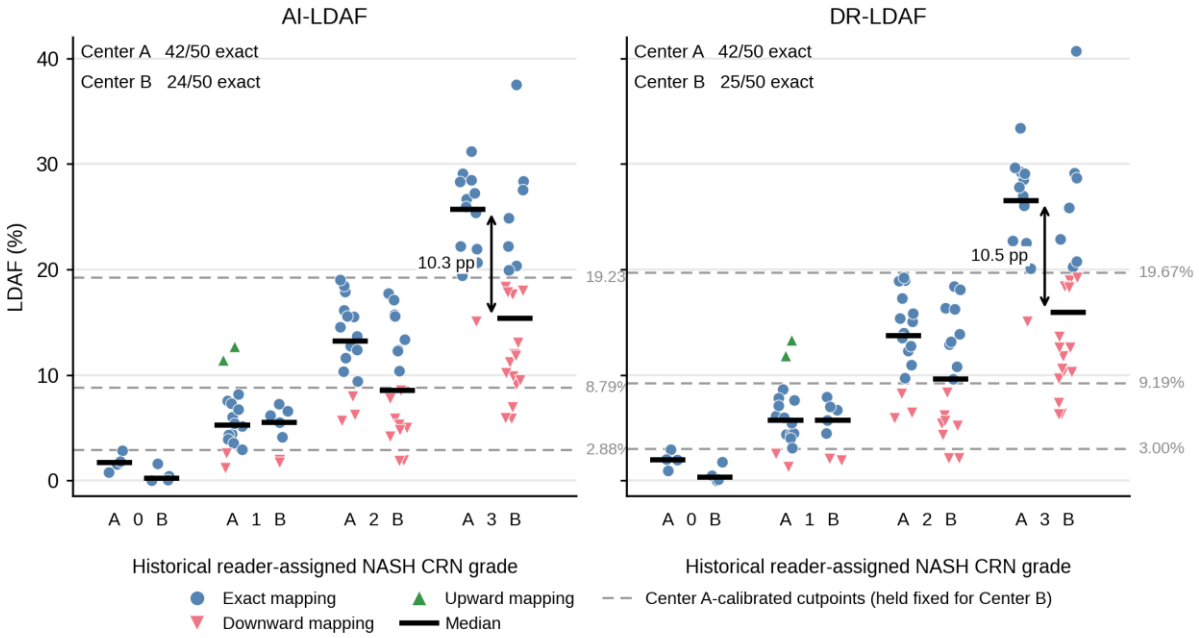


Fig. 2. Cross-center behavior of locked LDAF-to-historical-grade mappings. Case-level AI-LDAF (left) and expert-derived DR-LDAF (right; same-object control) are plotted against historical reader-assigned NASH CRN steatosis grade, with Center A (calibration) and Center B (locked application) shown side by side within each grade. Black bars mark medians; dashed lines mark the Center A-calibrated cutpoints held fixed for Center B; marker colors and symbols indicate exact, downward, and upward mappings. Panel counts show apparent in-sample Center A mappings; arrows show the between-center median differences at historical grade 3.

4. Discussion

Principal Findings. Close cross-center measurement agreement did not establish transportability of the downstream ordinal mapping. AI-LDAF and DR-LDAF agreed closely overall and within each center (CCC, 0.9968 overall; 0.9973 at Center A; 0.9961 at Center B), whereas the Center A-calibrated AI-LDAF mapping achieved only 48.0% exact agreement at Center B, compared with 76.6% optimism-corrected agreement at Center A. Repeating the analysis with DR-LDAF, which shared neither the lipid-droplet numerator nor the evaluable-tissue denominator with AI-LDAF, produced a parallel external decline, suggesting that AI-specific segmentation error alone did not readily explain the loss of agreement. All Center B discordances were downward and concentrated mainly among historical grades 2 and 3, for which Center B specimens had descriptively lower area fractions than Center A specimens assigned the same historical grade. Thus, the continuous measurement remained closely aligned with its same-object expert reference at both centers, whereas its mapping to historical grade did not transport. The observed cross-center limitation therefore concerned the LDAF-to-historical-grade relationship rather than the native continuous measurement itself.

Interpretation of the Measurement and Grade Dissociation. Although 36 cases carried historical grade 3, neither AI-LDAF nor DR-LDAF exceeded the nominal 66% grade-3 boundary. This does not imply that the historical grades were incorrect; rather, visually estimated hepatocyte involvement and computationally measured lipid-droplet area fraction represent different measurement constructs. NASH CRN visual grading thresholds therefore should not be applied directly as computational area-fraction cutoffs.

Close Measurement Agreement Did Not Ensure a Transportable Grade Mapping.

(1) Parallel external decline in AI-LDAF and DR-LDAF mappings. Exact agreement was 76.6% at Center A after optimism correction and 48.0% at Center B under locked application, with corresponding DR-LDAF values of 77.6% and 50.0%. Because DR-LDAF shared neither numerator nor denominator with AI-LDAF, the parallel decline was not readily explained by AI segmentation alone. Historical grades were non-standardized routine records, distinct from the prespecified reference implementation needed for future categorical models.

(2) **Continuous measurement from the analyzed study section.** X_i^{study} denotes the digitized H&E section analyzed for case i in this study, L_i denotes its continuous LDAF value, and f_{meas} denotes the measurement function. Their relationship is represented as follows:

$$L_i = f_{meas}(X_i^{study}) \tag{4}$$

(3) **Historical grade from an unreconstructed routine process.** The historical grade recorded for each case can instead be represented conceptually as:

$$G_i^{hist} = f_{hist}(H_i^{hist}; \mathcal{P}_i^{routine}), \quad G_i^{hist} \in \{0, 1, 2, 3\} \tag{5}$$

where H_i^{hist} denotes the histologic material used for routine grading, and $\mathcal{P}_i^{routine}$ denotes the reader- and workflow-related conditions under which the grade was assigned. Each historical grade record was linked to the same biopsy specimen as the corresponding study WSI. Correspondence to the exact H&E slide or histologic section could not be verified. The routine grading process was not reconstructed.

(4) **Scalar mapping based on L_i alone.** The mapping evaluated here was:

$$\hat{G}_i^{hist} = g(L_i) \tag{6}$$

where $g(\cdot)$ is the cutpoint rule calibrated in Center A and held fixed when applied to Center B, and $\hat{G}_i^{hist}$ is the projected historical grade. Because g uses L_i alone, it does not directly

incorporate information represented by H_i^{hist} and $\mathcal{P}_i^{routine}$. This decomposition provides a structural account of how close measurement agreement at both centers could coexist with failure of a locked mapping to transport. The present design does not identify which mechanism caused the decline. More broadly, validation of an upstream output need not carry over to a downstream projection. This study illustrates that distinction for a single measurement–grade pairing.

(5) **Construct-aligned reference for future categorical models.** Future work should establish a study-specific consensus reference grade for the existing NASH CRN steatosis construct on the exact H&E section analyzed, using independent review by trained and calibrated liver pathologists under a prespecified protocol with predefined consensus or adjudication. A CRN-compatible categorical model could then be developed from construct-aligned inputs, potentially including LDAF and additional morphologic features, and locked before external validation under the same reference protocol. This approach would standardize the reference implementation rather than redefine the NASH CRN categories.

Clinical Implications. These findings support further evaluation of AI-LDAF as a continuous whole-slide measurement with a visual overlay for pathologist review, but not for automated NASH CRN grading. Any future CRN-compatible categorical output should be developed against a prespecified reference standard established on the exact H&E section used for quantitative analysis. It should then be locked and externally validated using construct-aligned inputs that may include features beyond LDAF.[3,10] Prospective reader studies should assess whether AI-LDAF and its visual overlay improve pathologist reproducibility, efficiency, or clinical decision-making.

Limitations. This retrospective study included 100 cases from two Taiwanese centers that used the same scanner model. Cross-scanner reproducibility was assessed separately. The conservative 3/3 unanimity-intersection reference was not tested against alternative consensus rules. Because the three experts independently finalized the lipid-droplet masks from a shared medical-student pre-annotation rather than annotating de novo, shared omissions or anchoring may have influenced those masks. The ICC estimates therefore reflect agreement within this expert-finalization workflow. Historical grades were routine biopsy-level assessments linked to the same biopsy specimens from which the study WSIs were generated. The number and identity of readers may have varied across cases under each center's routine diagnostic workflow. Exact slide or section correspondence could not be verified, and no study-specific multi-reader consensus grading was performed. Reader-related, workflow-related, morphologic, case-mix, and center effects therefore could not be separated. The 50-case-per-center mapping analysis was exploratory and remained subject to cutpoint uncertainty despite optimism correction. Confidence intervals were not calculated for the optimism-corrected Center A estimates. Clinical outcomes, treatment response, action thresholds, and pathologist performance were not assessed.

5. Conclusion

A locked AI-derived lipid-droplet area measurement remained closely concordant with its same-object expert reference across centers, whereas its mapping to routine historical NASH CRN grades did not transport, a pattern reproduced with DR-LDAF. This study empirically shows that a transportable native measurement does not ensure a transportable downstream ordinal projection; the two should therefore be validated separately, with evidence matched to the claimed output and intended use. For AI-LDAF, the present evidence supports further evaluation as a continuous measurement with a visual overlay for pathologist review, but not as a basis for automated NASH CRN grading. Future CRN-compatible models should use prespecified reference standards established on the same H&E section and be locked before external validation.

References

1. Li, M., *et al.* Comparing morphometric, biochemical, and visual measurements of macrovesicular steatosis of liver. *Human Pathology* **42**, 356-360 (2011).
2. Franzén, L.E., Ekstedt, M., Kechagias, S. & Bodin, L. Semiquantitative evaluation overestimates the degree of steatosis in liver biopsies: a comparison to stereological point counting. *Modern Pathology* **18**, 912-916 (2005).
3. Pulaski, H., *et al.* Clinical validation of an AI-based pathology tool for scoring of metabolic dysfunction-associated steatohepatitis. *Nature Medicine* **31**, 315-322 (2025).
4. Davison, B.A., *et al.* Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. *Journal of Hepatology* **73**, 1322-1332 (2020).
5. Kleiner, D.E., *et al.* Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313-1321 (2005).
6. Takahashi, Y. & Fukusato, T. Histopathology of nonalcoholic fatty liver disease/nonalcoholic steatohepatitis. *World Journal of Gastroenterology: WJG* **20**, 15539–15548 (2014).
7. Roy, M., *et al.* Deep-learning-based accurate hepatic steatosis quantification for histological assessment of liver biopsies. *Laboratory Investigation* **100**, 1367-1383 (2020).
8. Wu, T.-H., *et al.* BSMatch: boundary segmentation and matching for lipid droplet quantification in diagnosis of non-alcoholic fatty liver disease. *IEEE Journal of Biomedical and Health Informatics* 29, 5912–5921 (2025).
9. Hall, A.R., *et al.* Hepatic steatosis estimated microscopically versus digital image analysis. *Liver International* **33**, 926-935 (2013).
10. Iyer, J.S., *et al.* AI-based automation of enrollment criteria and endpoint assessment in clinical trials in liver diseases. *Nature Medicine* **30**, 2914-2923 (2024).
11. Fernandez, A.I., *et al.* Examination of low ERBB2 protein expression in breast cancer tissue. JAMA Oncol 8, 607–610 (2022).
12. Bui, M.M., *et al.* Quantitative image analysis of human epidermal growth factor receptor 2 immunohistochemistry for breast cancer: guideline from the College of American Pathologists. *Archives of Pathology & Laboratory Medicine* **143**, 1180-1195 (2019).
13. U.S. Food and Drug Administration. Artificial Intelligence-Enabled Device Software Functions: Lifecycle Management and Marketing Submission Recommendations. Draft Guidance for Industry and Food and Drug Administration Staff (January 2025).
14. Filozof, C.M., *et al.* Best practices in liver biopsy histologic assessment for nonalcoholic steatohepatitis clinical trials: expert opinion. *GastroHep* **2022**, 3538103 (2022).
15. Shrout, P.E. & Fleiss, J.L. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin* **86**, 420-428 (1979).
16. Koo, T.K. & Li, M.Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine* **15**, 155-163 (2016).
17. Bulten, W., *et al.* Automated deep-learning system for Gleason grading of prostate cancer using biopsies: a diagnostic study. *The Lancet Oncology* **21**, 233-241 (2020).
18. Berman, M., Rannen Triki, A. & Blaschko, M.B. The Lovász-Softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4413–4421 (2018).

19. Willmott, C.J. & Matsuura, K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research* **30**, 79-82 (2005).
20. Chai, T. & Draxler, R.R. Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geoscientific Model Development* **7**, 1247–1250 (2014).
21. Bland, J.M. & Altman, D.G. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet* **1**(8476), 307–310 (1986).
22. Lin, L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255-268 (1989).
23. Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin* **70**, 213–220 (1968).
24. Harrell, F.E., Jr., Lee, K.L. & Mark, D.B. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* **15**, 361-387 (1996).
25. Steyerberg, E.W., Bleeker, S.E., Moll, H.A., Grobbee, D.E. & Moons, K.G. Internal and external validation of predictive models: a simulation study of bias and precision in small samples. *Journal of Clinical Epidemiology* **56**, 441-447 (2003).
26. Noma, H., Shinozaki, T., Iba, K., Teramukai, S. & Furukawa, T.A. Confidence intervals of prediction accuracy measures for multivariable prediction models based on the bootstrap-based optimism correction methods. *Statistics in Medicine* **40**, 5691-5701 (2021).