Title

Clinical validation of an AI-based pathology tool for scoring of metabolic dysfunction-associated steatohepatitis.

Permalink

https://escholarship.org/uc/item/9q87103h

Journal

Nature Medicine, 31(1)

Authors

Pulaski, Hanna
Harrison, Stephen
Mehta, Shraddha
et al.

Publication Date

2025

DOI

10.1038/s41591-024-03301-2

Peer reviewed

nature medicine

Article

https://doi.org/10.1038/s41591-024-03301-2

Clinical validation of an AI-based pathology tool for scoring of metabolic dysfunction-associated steatohepatitis

Received: 6 May 2024

Accepted: 16 September 2024

Published online: 4 November 2024

Check for updates

Hanna Pulaski[1,19], Stephen A. Harrison[2,19], Shraddha S. Mehta[1,12], Arun J. Sanyal[3], Marlena C. Vitali[1,13], Laryssa C. Manigat[1], Hypatia Hou[1], Susan P. Madasu Christudoss[1,14], Sara M. Hoffman[1,15], Adam Stanford-Moore[1], Robert Egger[1], Jonathan Glickman[1,16], Murray Resnick[1,17], Neel Patel[1], Cristin E. Taylor[1], Robert P. Myers[4], Chuhan Chung[5], Scott D. Patterson[6], Anne-Sophie Sejling[7], Anne Minnich[8], Vipul Baxi[8], G. Mani Subramaniam[4], Quentin M. Anstee[9], Rohit Loomba[10], Vlad Ratziu[11], Michael C. Montalto[1,18], Nick P. Anderson[1], Andrew H. Beck[1] & Katy E. Wack[1] ✉

Metabolic dysfunction-associated steatohepatitis (MASH) is a major cause of liver-related morbidity and mortality, yet treatment options are limited. Manual scoring of liver biopsies, currently the gold standard for clinical trial enrollment and endpoint assessment, suffers from high reader variability. This study represents the most comprehensive multisite analytical and clinical validation of an artificial intelligence (AI)-based pathology system, AI-based measurement of metabolic dysfunction-associated steatohepatitis (AIM-MASH), to assist pathologists in MASH trial histology scoring. AIM-MASH demonstrated high repeatability and reproducibility compared to manual scoring. AIM-MASH-assisted reads by expert MASH pathologists were superior to unassisted reads in accurately assessing inflammation, ballooning, MAS ≥ 4 with ≥1 in each score category and MASH resolution, while maintaining non-inferiority in steatosis and fibrosis assessment. These findings suggest that AIM-MASH could mitigate reader variability, providing a more reliable assessment of therapeutics in MASH clinical trials.

Metabolic dysfunction-associated steatotic liver disease (MASLD)[1] is emerging as an important global health challenge, affecting approximately a quarter of the global population[2]. The progression of MASLD to MASH has emerged as the foremost reason for liver transplants among women[3], with predictions suggesting that it may soon account for the leading overall cause of liver transplant[4]. The urgency of the situation is underscored by the limited number of approved therapeutic interventions by the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) for MASH, even though it affects a substantial number of patients worldwide. The landscape of drug development in this domain is fraught with trials that have shown borderline results or outright failures based on liver histology.

The challenge is exacerbated by the absence of a reliable and validated histologic scoring mechanism to ascertain patient suitability for clinical trials and to evaluate the success of experimental treatments. Histologic-based assessment of liver biopsies is currently the gold standard for MASH diagnosis. This diagnosis is based on the presence of specific histologic patterns observed in the absence of substantial alcohol consumption, and the patterns with extent of fibrosis play a pivotal role in disease staging. The FDA has recognized that alterations in

A full list of affiliations appears at the end of the paper. ✉e-mail: katy.wack@pathai.com

these histologic attributes, observable through liver biopsies, are likely indicative of clinical benefits[5]. Consequently, these score-based disease activity and stage changes are deemed viable surrogate endpoints in MASH clinical trials for accelerated approvals[6,7]. Key instruments like the MASLD activity score (MAS) by the MASH Clinical Research Network (CRN) facilitate disease activity measurement[8], while the CRN fibrosis scale evaluates fibrosis progression or improvement and is an influential predictor of long-term outcomes[9]. Regulatory bodies such as the FDA and the EMA predominantly rely on the CRN MASH measurement systems to determine surrogate endpoints[6,7,10].

The recent emergence of noninvasive tests (NITs) has led to an initiative to replace biopsies with NITs and has been discussed in the MASH community. However, there are no NITs or combination of NITs that are currently analytically or clinically validated for broad use in trials, which demonstrate high sensitivity, specificity and reproducible grading and staging of patients with MASH for use as surrogate endpoints in MASH clinical trials. Acquiring this validation data, including clinical outcomes across multiple drug candidates, will take years. Many biopsy-based MASH clinical trials are currently in phase 2 and 3 trials, and the recent accelerated approval of resmetirom was achieved through consensus scoring, with re-read methods used to confirm histologic score-based primary endpoints[11]. This burdensome approach can be necessary to overcome questions around reader bias and variability, which can affect the accuracy of histologic-based score change and, therefore, determination of whether a drug candidate has met its primary endpoint or to measure and monitor its efficacy. Additionally, the current gold-standard approach is still subject to substantial interpanel variability, demonstrating that there remains a lack of standardization and therefore of reliable, accurate scoring[12]. Full approval will not occur until clinical outcomes show a favorable benefit-to-risk profile in treated patients when these results are collected over multiple years. This substantial read variability is a major risk for potentially effective treatments to fail in phase 2b trials, which have relatively low sample sizes, and phase 3 trials or to require very costly and burdensome, multiple-read strategies to confirm and measure efficacy. Therefore, there is still an urgent unmet need for a tool that can be used by pathologists to enroll and measure histologic change for accelerated approval accurately, precisely and in a standardized manner. In addition, it will be important to understand the relationship between histologic-based assessments in validating NITs for diagnostic contexts of use.

The interpretation of the current scoring systems presents substantial challenges to clinical trial outcome analysis, particularly concerning reproducibility[13–17]. Given that the gold-standard endpoint for accelerated approval is a difference in histological scores from baseline to treatment time points, inherent intra-reader and inter-reader variability can confound the measurement of true drug effect[12]. This variability can substantially undermine the power of a study, posing challenges especially in trials with smaller sample sizes, such as phase 1 and 2 trials. To circumvent this limitation, trials are often required to be overpowered, adding cost and time to trials. Such variability likely arises due to discrepancies in feature interpretation, feature heterogeneity within a biopsy sample and the quantification of these features using scoring systems[18]. Additionally, the current scoring criteria were not developed to quantify changes in disease activity.

The rapidly evolving field of AI offers a promising avenue to address these challenges. AI has demonstrated notable advancements in numerous medical disciplines, with a marked increase in CE-marked (a standard for European health, safety, performance and environmental requirements) and FDA-approved in vitro diagnostics for AI-based medical devices and algorithms from 2015 to 2020 (ref. 19), including the FDA authorization for an AI product in digital pathology in 2021 for Paige Prostate[20]. However, the field of quantitative pathology in MASH therapeutic development still awaits a tool that is scalable, reproducible and validated. Recently, we described the development and verification of the AIM-MASH (AI-based measurement of nonalcoholic steatohepatitis) AI-based clinical trial tool[21]. In this previous body of work, the algorithm was developed and verified for accuracy compared to a panel of manual readers to confirm that the tool was ready to be locked. As proof of concept, the algorithm alone (without pathologist review) was also retrospectively deployed on the ATLAS clinical trial dataset to demonstrate the utility of the tool. The work presented here represents extensive, multisite analytical and clinical validation of the algorithm alone and as an assist to MASH pathologists, as it would be used prospectively in a clinical trial, with each histologic component score being assessed individually and as a part of histologic-based composite inclusion criteria and endpoint determination. This validation study, the largest known of its kind, included approximately 13,000 independent reads for over 1,400 biopsies across four completed, global MASH clinical trials with various drug mechanisms of action. The study was performed across multiple sites and included samples with extensive variation in disease activity as well as biopsy, staining and scanning quality. Multiple prospectively collected pathologist reads per case (in which readers were either unassisted (independent manual readers; IMR) or assisted by AI) were collected from MASH expert pathologists, including reads from an independent 'gold-standard' consensus group or ground truth (GT). These reads were used to externally and robustly test both the algorithm alone and as used as an aid to pathologists (Fig. 1a) in representative trial settings. This extensive collection of AIM-MASH validation studies and analyses was designed in partnership with the FDA, the EMA and multiple experts from academia and drug development over several years of collaborative work. The aim was to demonstrate the tool's ability to provide a reliable, efficient solution for pathologists to address the urgent unmet need for accurate, standardized, clinical trial enrollment and histologic endpoint assessments, paving the way for more streamlined MASH drug approval pathways.

Once a tool is analytically and clinically validated and is fully qualified by the FDA and the EMA in the Drug Development Tool and Novel Methodologies for Drug Development programs, it is then more broadly available for use by pathologists in place of manual scoring for all histologic assessments in MASH trials.

A clinical diagnostic intended use could require further training and/or validation to align with the clinical MASLD intended use population because this may differ from the clinical trial population but could be beneficial to pathologists in the diagnostic setting.

Results

Overlay validation analyses

The overlay validation was a substudy, independent of the analytical and clinical validations, designed to validate the use of the algorithm-generated overlays to assist the pathologist in reviewing the slide and AIM-MASH scores. Up to 160 frames or regions of interest within the whole-slide image (WSI) with a predefined area per feature (steatosis, lobular inflammation, hepatocellular ballooning, fibrosis, hematoxylin and eosin (H&E) artifact and trichrome artifact) were evaluated in this study (some frames were enrolled for multiple features). Distributions of frames based on slide-level score (GT scores) are listed in Extended Data Table 1, and distributions of frames based on frame-level scores (collected from the enrollment pathologist) are listed in Extended Data Table 2. For each frame and each feature, the pathologists indicated whether the feature was present (yes or no), shown in Extended Data Table 3.

The acceptance criteria for true positive (TP; evaluation of underestimation by overlay) success were met for all feature overlays except for hepatocellular ballooning, where it was narrowly missed, and the mean success rates were all above 0.85. H&E artifact TP success rate was 0.97 (95% confidence interval (CI), 0.95–0.99), trichrome artifact was 0.99 (95% CI, 0.97–1), lobular inflammation was 0.94 (95% CI, 0.92–0.96), steatosis was 0.96 (95% CI, 0.93–0.98), and fibrosis was

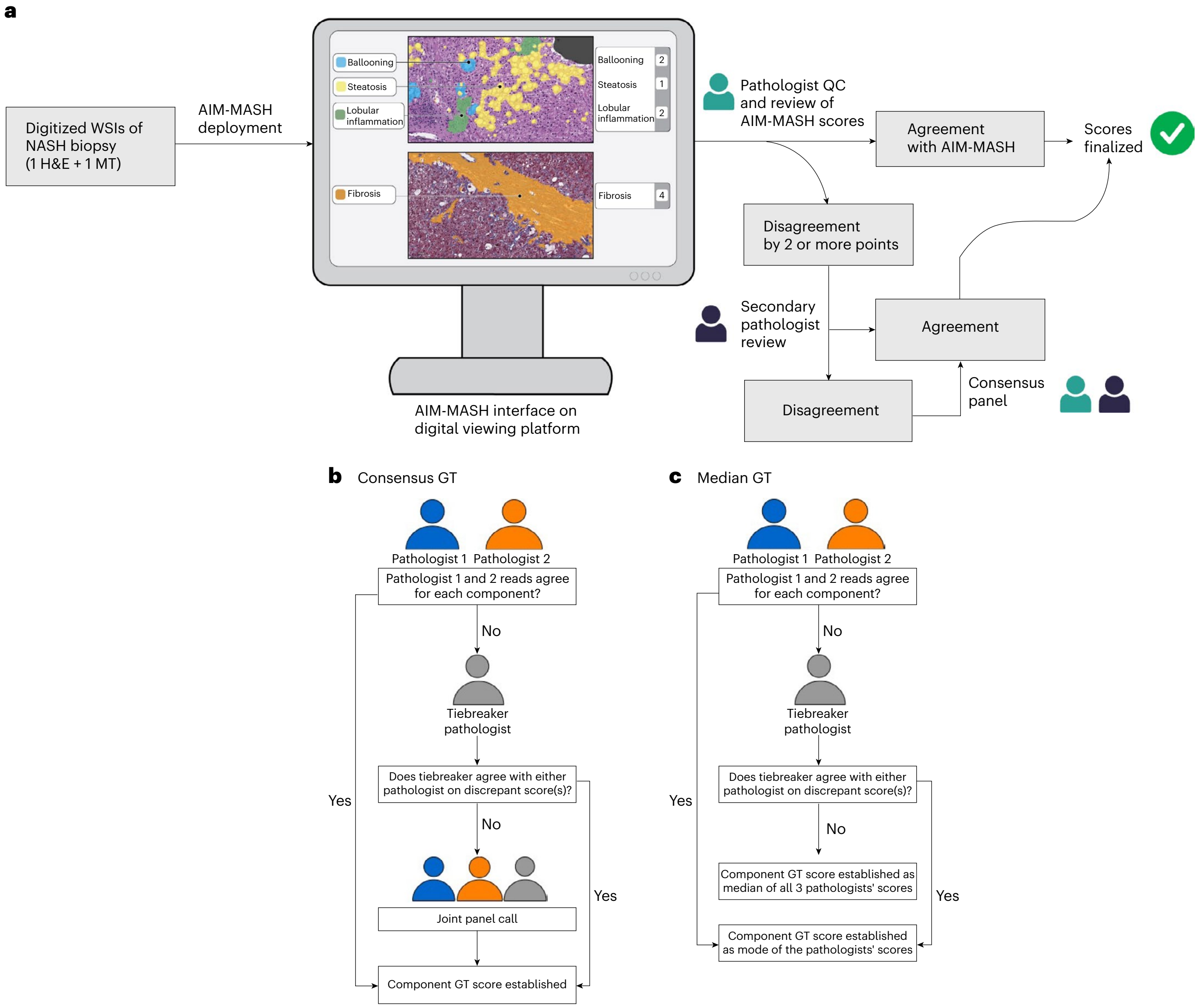

a
Digitized WSIs of NASH biopsy (1 H&E + 1 MT)
AIM-MASH deployment
Ballooning
Steatosis
Lobular inflammation
Fibrosis
Ballooning 2
Steatosis 1
Lobular inflammation 2
Fibrosis 4
AIM-MASH interface on digital viewing platform
Pathologist QC and review of AIM-MASH scores
Agreement with AIM-MASH
Scores finalized
Disagreement by 2 or more points
Secondary pathologist review
Agreement
Disagreement
Consensus panel
b
Consensus GT
Pathologist 1
Pathologist 2
Pathologist 1 and 2 reads agree for each component?
No
Yes
Tiebreaker pathologist
Does tiebreaker agree with either pathologist on discrepant score(s)?
No
Yes
Joint panel call
Component GT score established
c
Median GT
Pathologist 1
Pathologist 2
Pathologist 1 and 2 reads agree for each component?
No
Yes
Tiebreaker pathologist
Does tiebreaker agree with either pathologist on discrepant score(s)?
No
Yes
Component GT score established as median of all 3 pathologists' scores
Component GT score established as mode of the pathologists' scores

Fig. 1 | AI-assisted workflow with representative AIM-MASH overlays and GT panel workflows. a, In the AI-assisted workflow, the primary pathologist reviews the AIM-MASH output and does a quality control (QC) review of the Hematoxylin and Eosin (H&E) and Masson's Trichrome (MT) slides (determines whether restaining or rescanning of the slide is necessary, confirms that all trial-specific criteria are met and notes any additional findings). If the primary pathologist disagrees with any MASH component(s) by two points or more, the case goes to a review by a secondary pathologist, who independently reviews the discordant AIM-MASH score(s). If the secondary pathologist agrees with the primary pathologist's modified score, this will be the final score; if they disagree with the primary pathologist or agree with AIM-MASH, the two pathologists will convene on a consensus call in which they agree on the final score. **b**, Consensus GT for each biopsy was determined by one of two panels of hepatopathologists. Each panel consisted of two main reader pathologists and an auxiliary tiebreaker pathologist. Discrepancies in scoring among the primary readers prompted the intervention of the tiebreaker pathologist, who was blind to initial assessments. When the tiebreaker's scoring diverged from that of both primary readers, a panel discussion was convened for consensus, with the tiebreaker's score being decisive in rare cases of continued disagreement. **c**, For the median GT score, when the tiebreaker's scoring diverged from that of both primary readers, the median of the three scores was considered final. Overall, five distinct pathologists contributed to establishing the GT.

0.97 (95% CI, 0.95–0.99). For hepatocellular ballooning, the overall TP success rate was 0.87, with 95% CI (0.83–0.91). The acceptance criteria for the false positive (FP; evaluation of overestimation by overlay) success rate was met for all six feature overlays. H&E artifact success rate for FP was 0.97 (95% CI, 0.95–0.99), trichrome artifact was 0.93 (95% CI, 0.90–0.96), lobular inflammation was 0.99 (95% CI, 0.98–0.99), steatosis was 1.00 (95% CI, 0.98–1), hepatocellular ballooning was 0.92 (95% CI, 0.90–0.94), and fibrosis was 0.99 (95% CI, 0.99–1).

The individual pathologist TP and FP success rates are listed in Table 1. The number of frames for which all three evaluating pathologists agreed on the presence of the feature (independent of any overlay) divided by the number of frames for which at least one pathologist indicated the presence of feature in a frame was 89% (132 of 148 frames) for H&E artifact, 55.1% for hepatocellular ballooning (65 of 118 frames), 80.0% (124 of 155 frames) for lobular inflammation, 99.4% (158 out of 159 frames) for steatosis, 72.0% (108 of 150 frames) for trichrome artifact and 96.8% (149 of 154 frames) for fibrosis. Given that the agreement for the presence of hepatocellular ballooning was the lowest (55.1%) and the TP success rate for ballooning was above 0.90 for two of three of the pathologists, the sources of variability between pathologists for this feature were further examined. For the 65 frames for which all three evaluating pathologists indicated the presence of hepatocellular

Table 1 | TP and FP success rates per individual pathologist for overlay validation

Feature	Pathologist	TP success rate (95% CI)	FP success rate (95% CI)
H&E artifact	A	0.97 (0.94, 0.99)	0.96 (0.92, 0.99)
	B	0.98 (0.95, 1.0)	0.98 (0.95, 0.99)
	C	0.97 (0.94, 0.99)	0.99 (0.97, 1.0)
Hepatocellular ballooning	A	0.96 (0.91, 0.99)	1.00 (0.98, 1.0)
	B	0.94 (0.89, 0.99)	1.00 (0.98, 1.0)
	C	0.72 (0.64, 0.81)	0.76 (0.70, 0.83)
Lobular inflammation	A	0.98 (0.95, 1.0)	1.00 (0.98, 1.0)
	B	0.98 (0.95, 1.0)	1.00 (0.98, 1.0)
	C	0.86 (0.81, 0.92)	0.98 (0.95, 0.99)
Steatosis	A	0.94 (0.90, 0.98)	1.00 (0.98, 1.0)
	B	0.99 (0.97, 1.0)	1.00 (0.98, 1.0)
	C	0.94 (0.90, 0.98)	1.00 (0.98, 1.0)
Trichrome artifact	A	0.99 (0.97, 1.0)	0.88 (0.82, 0.92)
	B	0.98 (0.96, 1.0)	0.94 (0.91, 0.98)
	C	0.99 (0.97, 1.0)	0.98 (0.95, 0.99)
Fibrosis	A	0.97 (0.94, 0.99)	1.00 (0.98, 1.0)
	B	0.99 (0.98, 1.0)	1.00 (0.98, 1.0)
	C	0.95 (0.91, 0.98)	0.99 (0.98, 1.0)

ballooning, the TP success rate was calculated. Pathologists A and B identified underestimation in one and three of the 65 frames, respectively, resulting in TP success rates of 0.99 for pathologist A and 0.95 for pathologist B for those frames. However, pathologist C identified underestimation in ten of the 65 frames, showing a TP success rate of 0.85. Additionally, pathologist C identified a total of 111 frames that had some ballooned cells compared to 92 and 71 for pathologists A and B (Extended Data Table 3), indicating that pathologist C may identify more cells as ballooned hepatocytes than the other two pathologists and the algorithm. This is predictable given the lack of standardization across expert pathologists in both identifying and quantifying ballooned hepatocytes[22].

Algorithm repeatability and reproducibility

For interday scanner repeatability (AIM-MASH deployment on the same glass slides on different scans from the same scanner on different days), the mean agreement rate between the AIM-MASH scoring on the three separate WSIs for steatosis was 0.93 (95% CI, 0.89–0.96; $P < 0.0001$), for lobular inflammation was 0.96 (95% CI, 0.94–0.99; $P < 0.0001$), for hepatocellular ballooning was 0.96 (95% CI, 0.93–0.98; $P < 0.0001$) and for fibrosis was 0.93 (95% CI, 0.89–0.96; $P < 0.001$) (Fig. 2a).

For intersite scanner reproducibility (AIM-MASH deployment on the same glass slides on different scans from three different sites), the mean agreement rate for hepatocellular ballooning was 0.91 (95% CI, 0.87–0.95; $P = 0.02$), meeting the acceptance criteria. The mean agreement rates for steatosis, lobular inflammation and fibrosis were approximately 85%, but the CIs fell slightly below the 0.85 acceptance criteria (steatosis, 0.86 (95% CI, 0.81–0.9; $P = 0.39$); lobular inflammation, 0.85 (95% CI, 0.80–0.89; $P = 0.53$); fibrosis, 0.87 (95% CI, 0.82–0.91; $P = 0.21$)) (Fig. 2b).

Pairwise inter-reader agreements were calculated between IMR pathologists across all cases (Supplementary Table 1) to explicitly compare reproducibility across study pathologists to reproducibility achieved by AIM-MASH across sites and scanners. For all histologic components, interscan, intrasite repeatability and interscan, intersite reproducibility were higher than for pathologist mean pairwise agreement (for pairs of pathologists who read at least ten common cases) (Table 2).

Accuracy of the algorithm alone and as a pathologist-assist tool

Evaluation of the non-inferior accuracy of AIM-MASH (algorithm only and AI assisted) to IMRs was assessed in 1,481 cases by comparing the mean weighted kappa (WK) of IMRs with GT (workflow in Fig. 1b) to the WK of AIM-MASH with GT (workflow in Fig. 1b (Fig. 3)).

For AIM-MASH only (Fig. 3a), the difference in WK for AIM-MASH and GT compared to mean WK for IMR and GT for hepatocellular ballooning was 0.15 (95% CI, 0.11–0.18; non-inferiority $P < 0.0001$) and for lobular inflammation was 0.12 (95% CI, 0.08–0.17; non-inferiority $P < 0.0001$) with $P < 0.0001$ for superiority for both components. The difference in WK for AIM-MASH only and GT compared to WK of mean IMR and GT for steatosis was 0.01 (95% CI, −0.02 to 0.03; non-inferiority $P < 0.0001$) and for fibrosis was −0.01 (95% CI, −0.04 to 0.02; non-inferiority $P < 0.0001$). Steatosis and fibrosis met non-inferiority but did not achieve superiority.

For AI-assisted pathologist reading of the 1,481 cases (Fig. 3b), the difference in WK for AI assisted and GT compared to mean WK for IMR and GT for hepatocellular ballooning was 0.15 (95% CI, 0.11–0.19; non-inferiority $P < 0.0001$) and for lobular inflammation was 0.12 (95% CI, 0.08–0.17; non-inferiority $P < 0.0001$) with $P < 0.0001$ for superiority for both components. The difference in WK for AI assisted and GT compared to mean WK for IMR and GT for steatosis was 0.01 (95% CI, −0.02 to 0.04; non-inferiority $P < 0.0001$) and for fibrosis was 0.01 (95% CI, −0.02 to 0.03; non-inferiority $P < 0.0001$). Steatosis and fibrosis met non-inferiority but did not achieve superiority. For all MASH score components, WKs for AI assisted and GT were in the ranges of published CRN pathologist WKs[8,14].

For AI-assisted pathologist reading, accuracy was higher for composite histologic scores than for IMRs (Fig. 3c). The WKs for AI assisted and GT and WKs for IMR and GT for fibrosis 2 and 3 (F2 and F3) versus other were equivalent, with WK for AI assisted and GT being slightly higher than WK for IMR and GT (0.57 versus 0.53, respectively; Fig. 3c). WKs for the trial-relevant enrollment criteria MAS ≥ 4 with ≥ 1 in each score category between AI assisted and GT were significantly (lower bound (LB) of the 95% CI for AI assisted versus GT kappa was greater than the upper bound of the 95% CI for IMR versus GT kappa) higher than the WK between IMR and GT (0.63 versus 0.51, respectively, with a difference of 0.11 and a 95% CI of 0.07–0.16) and, for MASH resolution (defined as a hepatocellular ballooning score of 0, a lobular inflammation score of 0 or 1 and any steatosis score) between AI assisted and GT, were also significantly higher than the WK between IMR and GT (0.54 versus 0.37, respectively, with a difference of 0.16 and a 95% CI of 0.10–0.22) (Fig. 3c).

For AI-assisted evaluation against a median of a panel of pathologists (GT workflow described in Fig. 1c), non-inferiority was met for all histologic components for agreement of AI-assisted reads with median GT reads, compared to the agreement between median read scores derived from two different groups of pathologists (GT workflow in Fig. 1c, results in Fig. 4). For steatosis, the average WK for AI assisted versus GT was 0.68 and for manual median versus GT was 0.75, with a difference of −0.07; for lobular inflammation, the WK for AI assisted versus GT was 0.43 and for manual median versus GT was 0.44, with a difference of −0.02; for hepatocellular ballooning, the WK for AI assisted versus GT was 0.56 and for manual median versus GT was 0.53, with a difference of 0.04; and, for fibrosis, the WK for AI assisted versus GT was 0.65 and for manual median versus GT was 0.72, with a difference of −0.09.

Discussion

AI-based tools have the potential to solve many issues around standardized, accurate and reproducible scoring, within and across trials. Multiple pathologists can assess biopsies on validated WSI viewers[23] for sample adequacy and evaluability and for overall diagnosis and

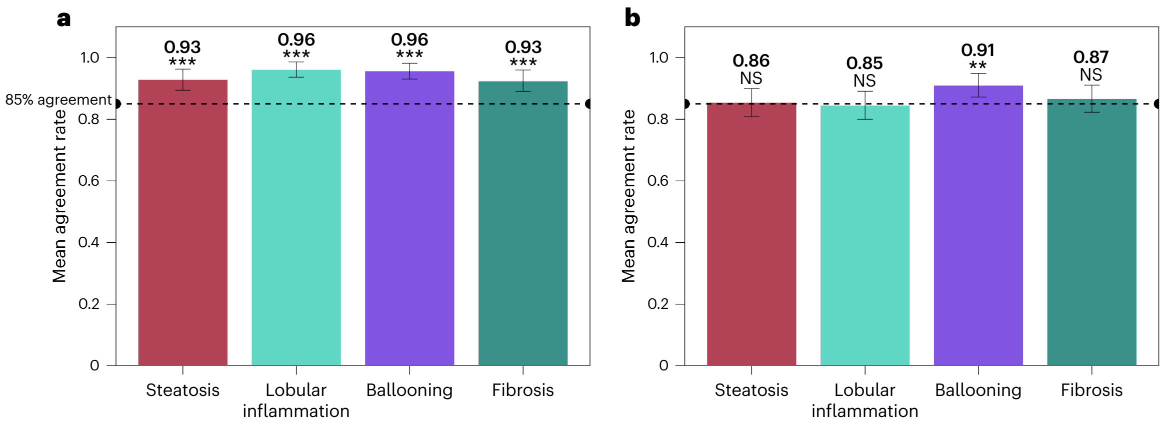


Fig. 2 | Scanner repeatability and reproducibility of AIM-MASH. a, For scanner repeatability, a subset of 150 cases from the clinical validation were scanned multiple times using the same Leica Aperio AT2 scanner at ×40 magnification on three nonconsecutive days (intrasite, interscan). **b**, For scanner reproducibility, the same slides were scanned once at three different laboratories by three different operators using three different Leica Aperio AT2 scanners at ×40 magnification (intersite). Bootstrap percentile *P* values showing statistical significance for the one-sided hypothesis that the mean agreement rate between algorithm scores for each scan is greater than 0.85 are as follows: ****P* < 0.0001; ***P* < 0.01; **P* < 0.05; not significant (NS), *P* ≥ 0.05. Whiskers show the 95% CIs for mean agreement rate estimated using 2,000 bootstraps. Dashed lines indicate 85% agreement.

Table 2 | Manual pathologist versus AIM-MASH repeatability and reproducibility

Feature	Mean AIM-MASH interscan, intrasite repeatability (% agreement)	Mean AIM-MASH intersite reproducibility (% agreement)	Mean pairwise agreement for pathologists (% agreement)
Steatosis	93.1	85.6	70.3
Lobular inflammation	95.8	84.7	45.3
Hepatocellular ballooning	96.3	91.2	55.6
Fibrosis	92.6	86.8	61.5

additional findings, while using AI tools to efficiently provide the accurate, standardized and consistent scores needed.

The AIM-MASH outputs have been validated according to their proposed use with representative trial datasets, including a sample of screen failures and a majority of enrolled patients at risk for MASH (defined as nonalcoholic fatty liver disease (NAFLD) activity score (NAS) ≥ 4 with fibrosis score ≥2), both baseline and follow-up time points variable in disease activity, stain, scanning site and drug candidate intervention. Furthermore, the overlays presented to the pathologist, identifying key areas that the model predicts as artifact, steatosis, hepatocellular ballooning, lobular inflammation and fibrosis, have been validated by multiple pathologist readers on a frame level, demonstrating that they are highly sensitive and sufficiently specific in playing their role as a highlighter to guide pathologist review, along with the associated model scores. The variability of overlay validation results observed between pathologists for ballooning was as expected[22] and highlights the need for an accurate, reproducible scoring tool, such as was demonstrated by AIM-MASH here (Figs. 2 and 3a), combined with a workflow that allows for diagnosis confirmation and sample and score quality control, but limits when the pathologist can change the algorithm score (the two-point rejection workflow used here) to maximize standardization and minimize individual bias. These results demonstrate the precision of AIM-MASH in measuring each component of the CRN scoring system in liver biopsies from patients screened and/or enrolled in a MASH clinical trial.

Repeatability studies demonstrated superior performance of AIM-MASH when compared to a performance goal of 85% as well as to relevant published manual intrapathologist trial read agreements (steatosis, 0.72; lobular inflammation, 0.55; hepatocellular ballooning, 0.70; fibrosis, 0.72) described in the literature[14]. AIM-MASH reproducibility across the three external laboratories, using different operators and different Leica Aperio AT2 scanners, was higher for all MASH components than published interpathologist variability across expert MASH pathologists (0.63 for steatosis, 0.60 for lobular inflammation, 0.63 for hepatocellular ballooning and 0.51 for fibrosis)[14]. Furthermore, the repeatability and reproducibility agreement achieved in this study with AIM-MASH was higher than the interpathologist agreement for IMRs.

Finally, the clinical validation study demonstrated that AIM-MASH consistently brought individual pathologists closer to GT reads (approach in Fig. 1b) for the histologic components historically most difficult to score (hepatocellular ballooning and lobular inflammation) while maintaining high levels of accuracy for steatosis and fibrosis. To evaluate AIM-MASH reads against a statistical consensus currently being used as a gold-standard read during MASH trials, the agreement of AI-assisted reads with the median consensus of the GT reads (approach in Fig. 1c), 'panel 1', was compared to the agreement between two different median consensus groups (derived from GT pathologist reads, 'panel 1' or IMR pathologist reads, 'panel 2') in the same non-inferiority analysis used in the primary endpoint for accuracy. AI-assisted reads achieved non-inferiority for every histologic component score in this analysis, and AI-assisted read agreement with median GT for hepatocellular ballooning was higher than that for median IMR agreement with median GT. For steatosis, although the two manual median groups' mean agreement with each other was higher than that for AIM-MASH versus median GT, AI-assisted reads were still within the non-inferiority margin. Additionally, accuracy and reproducibility are interconnected in the MASH trial context of use for assessment of primary endpoints, and AIM-MASH provides a more reliable, reproducible read across all components. Furthermore, the gold standard is still subject to enrollment bias and lack of standardization, as demonstrated by the kappas achieved by the median IMR versus the median GT in this study (Fig. 4) and supported by findings from Sanyal et al.[12], which evaluated agreement between two gold-standard panel reads. Finally, the achievement of non-inferiority by AIM-MASH for accuracy compared to a gold-standard read across a robust clinical validation dataset provides strong evidence that AIM-MASH agrees with two consensus groups (panel 1 and panel 2) as well as that they agree with each other and, therefore, could replace the current gold-standard consensus read approach for trials while enabling a more-standardized,

https://doi.org/10.1038/s41591-024-03301-2

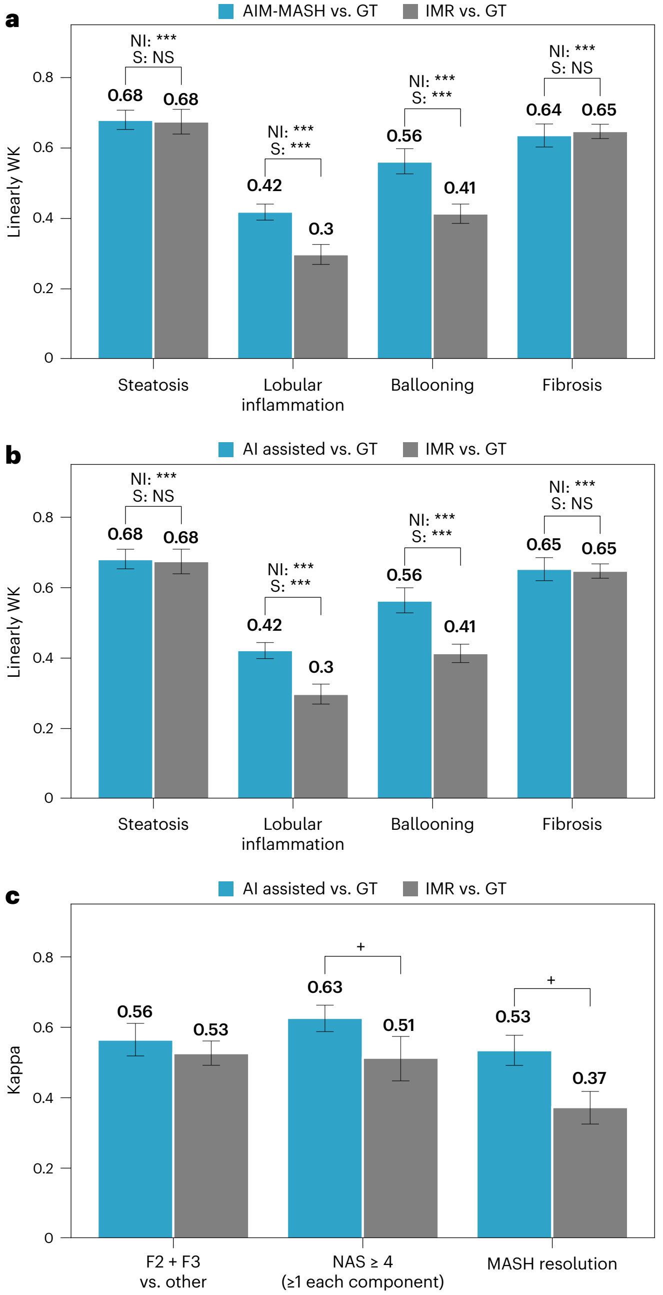


Fig. 3 | Accuracy concordance comparison of MASH histologic components and comparisons for MASH aggregate component scores (F2 and F3 versus other and NAS ≥ 4 with ≥1 in each score category versus other) and MASH resolution. a,**b**, Accuracy comparison, based on linearly WK, between AIM-MASH (without pathology review) versus GT and IMR versus GT in **a** and between AI assisted (AIM-MASH with pathology review) versus GT and IMR versus GT in **b** for MASH components. **c**, Accuracy comparison, based on kappa, between AI assisted versus GT and IMR versus GT for aggregate components relevant to clinical trial enrollment and endpoint criteria, including the score-based enrollment requirement, MAS ≥ 4 with a score of at least one for each component, fibrosis score of 2 or 3, and the NASH resolution endpoint, defined as a ballooning score of 0, a lobular inflammation score of 0 or 1 and any score for steatosis. Point estimates are shown on top of each bar, with whiskers representing the 95% CIs estimated from 2,000 bootstrap samples. Non-inferiority (NI) was assessed using bootstrap percentile *P* values for testing the one-sided hypothesis that the LB of the 95% CIs of the difference in AIM-NASH versus GT or AI assisted versus GT and IMR versus GT is not smaller than −0.1. S (superiority) was assessed by testing the one-sided hypothesis that the LB of the difference is greater than 0. ***$P < 0.0001$; **$P < 0.01$; *$P < 0.05$; NS, $P \geq 0.05$. '+' in **c** indicates aggregate components where the LB of the 95% CIs for AI assisted versus GT kappa is greater than the upper bound of the IMR versus GT kappa.

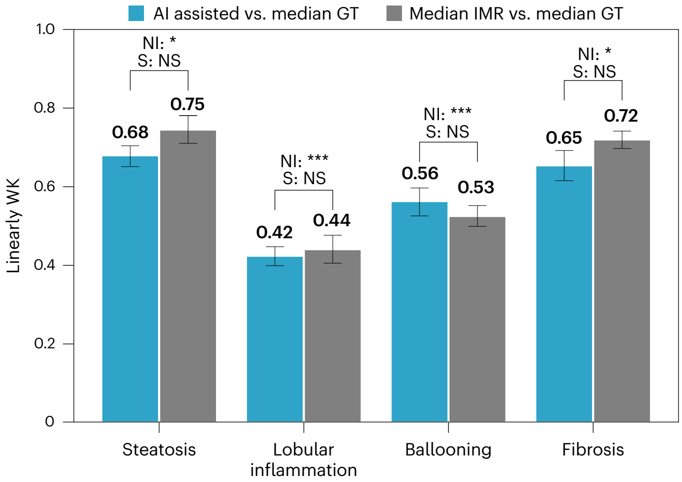


Fig. 4 | WK analysis for MASH components AI assisted and median panel comparisons. The same cohort of 1,481 cases used in analytical and clinical validation was used to determine the accuracy of AI-assisted reads against two panels of readers. Median GT (panel 1, using median scores, described in Fig. 1c), instead of panel calls for consensus and median IMR (panel 2), derived from a minimum of three IMRs, was determined. AI-assisted scores for each component met the non-inferiority performance criteria described in Statistical analysis (Methods). Superiority was not observed for any of the components. Whiskers represent 95% CIs estimated using 2,000 bootstrap samples. ***$P < 0.0001$; **$P < 0.01$; *$P < 0.05$; NS, $P \geq 0.05$.

less-biased approach to accurately enrolling and determining changes in score over time for primary histologic endpoints.

The sum of the ordinal scores for steatosis, lobular inflammation and hepatocellular ballooning (MAS) being greater than or equal to 4 (MAS ≥ 4) is one of the main indicators for a probable MASH diagnosis as well as commonly being a requirement for trial inclusion. Additionally, one component of the composite endpoint is MASH resolution, defined as a hepatocellular ballooning score of 0, a lobular inflammation score of 0 or 1 and any score for steatosis. AI-assisted reads for MAS ≥ 4 with ≥1 in each component category and for MASH resolution were superior compared to IMRs in their agreement with GT (Figs. 1b and 3c) This is an important indicator that AIM-MASH can be a powerful tool in increasing accuracy and standardizing key aspects of trial scoring for enrollment and for FDA- and EMA-recommended endpoints.

The AIM-MASH algorithm alone has also been demonstrated to either recapitulate or demonstrate that primary efficacy results were met across several trials and drug candidates (semaglutide, pegbelfermin, resmetirom[24–27]). In a phase 2b study for pegbelfermin, AIM-MASH revealed a statistically significant difference in the proportion of primary endpoint responders in treatment versus placebo groups, whereas the central pathologist scoring did not reveal a statistically significant difference[25]. In a phase 2b study for resmetirom, all endpoints met via both individual manual readers used in the trial were also met by AIM-MASH[26]. In the phase 3 study for resmetirom, for both MASH resolution and the fibrosis improvement endpoint, the percentages of patients who responded were comparable when assessed by AIM-MASH or manual pathology assessment[27]. Lastly, in a cirrhotic patient population from another phase 2 study for semaglutide, a numerically higher proportion of patients was seen across both assessment methods (AIM-MASH and manual reads) for semaglutide versus placebo for inflammation, steatosis and ballooning from baseline to week 48. Additionally, a lower placebo effect response was observed with AIM-MASH than with manual reads[24]. This supportive evidence, along with the accuracy of AIM-MASH alone and as an AI assist to pathologists, demonstrates the robust nature of AIM-MASH across a wide range of disease activity and in multiple phases of clinical drug trials.

As the samples for this study were sourced from completed clinical trials with a wide range of sample quality and the reads were performed retrospectively, the limitations of the study include the inability of the pathologists to request a restain or a rescan of samples when they thought the sample was not of sufficient quality. This could have led to higher rates of samples being deemed inadequate or non-evaluable for scoring, as, in a clinical trial setting, these samples could be restained or rescanned. However, these cases represented less than 4% of all clinical validation cases. Additionally, although the dataset was large and robust, new trial populations and/or drug candidates with novel mechanisms of actions not encountered here could potentially present a challenge to the algorithm in its current state. This highlights the importance of the pathologist evaluation and quality control of the algorithm results. Performance monitoring will be used to indicate when there may be room for future improvement through additional training.

Together, the above data support the use of AIM-MASH by pathologists in trials, and this use can play an important role in resolving the accuracy and precision gaps in MASH assessment, while guiding pathologists in an efficient evaluation to a standardized and reproducible score within and across trials. This in turn could substantially benefit patients with MASH by helping to bring truly effective therapies to market.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-024-03301-2.

References

1. Rinella, M. E. et al. A multisociety Delphi consensus statement on new fatty liver disease nomenclature. *J. Hepatol.* **79**, 1542–1556 (2023).
2. Younossi, Z. M. et al. Global epidemiology of nonalcoholic fatty liver disease—meta-analytic assessment of prevalence, incidence, and outcomes. *Hepatology* **64**, 73–84 (2016).
3. Noureddin, M. et al. NASH leading cause of liver transplant in women: updated analysis of indications for liver transplant and ethnic and gender variances. *Am. J. Gastroenterol.* **113**, 1649–1659 (2018).
4. Friedman, S. L., Neuschwander-Tetri, B. A., Rinella, M. & Sanyal, A. J. Mechanisms of NAFLD development and therapeutic strategies. *Nat. Med.* **24**, 908–922 (2018).
5. FDA–NIH Biomarker Working Group. *BEST (Biomarkers, Endpoints, and other Tools) Resource* (2016); https://www.ncbi.nlm.nih.gov/books/NBK326791/
6. Food and Drug Administration. *Nonalcoholic Steatohepatitis with Compensated Cirrhosis: Developing Drugs for Treatment Guidance for Industry — Draft* www.fda.gov/media/127738/download (2019).
7. Food and Drug Administration. *Noncirrhotic Nonalcoholic Steatohepatitis With Liver Fibrosis: Developing Drugs for Treatment Guidance for Industry* www.fda.gov/media/119044/download (2018).
8. Brunt, E. M., Kleiner, D. E., Wilson, L. A., Sanyal, A. J. & Neuschwander-Tetri, B. A. Improvements in histologic features and diagnosis associated with improvement in fibrosis in nonalcoholic steatohepatitis: results from the Nonalcoholic Steatohepatitis Clinical Research Network treatment trials. *Hepatology* **70**, 522–531 (2019).
9. Tong, X. F. et al. Histological assessment based on liver biopsy: the value and challenges in NASH drug development. *Acta Pharmacol. Sin.* **43**, 1200–1209 (2022).
10. European Medicines Agency. *Reflection Paper on Regulatory Requirements for the Development of Medicinal Products for Non-Alcoholic Steatohepatitis (NASH) (EMA/CHMP/111529/2024)* www.ema.europa.eu/en/documents/scientific-guideline/reflection-paper-regulatory-requirements-development-medicinal-products-non-alcoholic-steatohepatitis-nash_en.pdf (2023).
11. Harrison, S. A. et al. A phase 3, randomized, controlled trial of resmetirom in NASH with liver fibrosis. *N. Engl. J. Med.* **390**, 497–509 (2024).
12. Sanyal, A. J. et al. Utility of pathologist panels for achieving consensus in NASH histologic scoring in clinical trials: data from a phase 3 study. *Hepatol. Commun.* **8**, e0325 (2024).
13. Kleiner, D. E. et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. *Hepatology* **41**, 1313–1325 (2005).
14. Davison, B. A. et al. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. *J. Hepatol.* **73**, 1322–1332 (2020).
15. Merriman, R. B. et al. Correlation of paired liver biopsies in morbidly obese patients with suspected nonalcoholic fatty liver disease. *Hepatology* **44**, 874–880 (2006).
16. Juluri, R. et al. Generalizability of the Nonalcoholic Steatohepatitis Clinical Research Network histologic scoring system for nonalcoholic fatty liver disease. *J. Clin. Gastroenterol.* **45**, 55–58 (2011).
17. Pavlides, M. et al. Interobserver variability in histologic evaluation of liver fibrosis using categorical and quantitative scores. *Am. J. Clin. Pathol.* **147**, 364–369 (2017).
18. Harrison, S. A. et al. Insulin sensitizer MSDC-0602K in non-alcoholic steatohepatitis: a randomized, double-blind, placebo-controlled phase IIb study. *J. Hepatol.* **72**, 613–626 (2020).
19. Muehlematter, U. J., Daniore, P. & Vokinger, K. N. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. *Lancet Digit. Health* **3**, e195–e203 (2021).
20. Perincheri, S. et al. An independent assessment of an artificial intelligence system for prostate cancer detection shows strong diagnostic accuracy. *Mod. Pathol.* **34**, 1588–1595 (2021).
21. Iyer, J. S. et al. AI-based automation of enrollment criteria and endpoint assessment in clinical trials in liver diseases. *Nat. Med.* https://doi.org/10.1038/s41591-024-03172-7 (2024).
22. Brunt, E. M. et al. Complexity of ballooned hepatocyte feature recognition: defining a training atlas for artificial intelligence-based imaging in NAFLD. *J. Hepatol.* **76**, 1030–1041 (2022).
23. Pulaski, H. et al. Validation of whole slide image management system for metabolic-associated steatohepatitis for clinical trials. *J. Pathol. Clin. Res.* **10**, e12395 (2024); https://doi.org/10.1002/2056-4538.12395
24. Loomba, R. et al. Comparison of the effects of semaglutide on liver histology in patients with non-alcoholic steatohepatitis cirrhosis between machine learning model assessment and pathologist evaluation (Poster presentation). *2022 American Association for the Study of Liver Diseases* (*AASLD*). https://pathaiwp.wpenginepowered.com/wp-content/uploads/2023/01/FINAL_Loomba_AASLD_PathAI-key-results_poster_Approval-Draft_26Oct22.pdf (2022).
25. Shevell DE, Brown E, Du S, et al. Comparison of manual vs machine learning approaches to liver biopsy scoring for NASH and fibrosis: A post hoc analysis of the FALCON 1 study. *Hepatology* **74**, 1415A (2021).
26. Harrison, S. et al. Retrospective AI-based measurement of NASH histology (AIM-NASH) analysis of biopsies from Phase 2 study of Resmetirom confirms significant treatment-induced changes in histologic features of non-alcoholic steatohepatitis. *J. Hepatol.* **7**, S711–S712 (2022).

27. Iyer, J. et al. Artificial intelligence-based measurement of NASH histology (AIM-NASH) recapitulates primary results from phase 3 study of resmetirom for treatment of NASH/MASH. *Hepatology* **79**, E56–E57 (2023).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[1]PathAI, Inc., Boston, MA, USA. [2]Pinnacle Clinical Research, San Antonio, TX, USA. [3]Virginia Commonwealth University, Richmond, VA, USA. [4]OrsoBio, Inc., Palo Alto, CA, USA. [5]Inipharm, Inc., Bellevue, WA, USA. [6]Gilead Sciences, Inc., Foster City, CA, USA. [7]Novo Nordisk, Bagsvaerd, Denmark. [8]Bristol Myers Squibb, Princeton, NJ, USA. [9]Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University, Newcastle upon Tyne, UK. [10]UCSD School of Medicine, San Diego, CA, USA. [11]Sorbonne Université, ICAN Institute for Cardiometabolism and Nutrition, Assisstance Publique Hôpitaux de Paris, INSERM UMRS, Paris, France. [12]Present address: AbbVie, Inc., Irvine, CA, USA. [13]Present address: Whoop Inc., Boston, MA, USA. [14]Present address: Invicro, Needham, MA, USA. [15]Present address: Harvard Medical School, Boston, MA, USA. [16]Present address: Department of Pathology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA. [17]Present address: Department of Pathology and Laboratory Medicine, Rhode Island Hospital, Brown University, Providence, RI, USA. [18]Present address: Amgen, Thousand Oaks, CA, USA. [19]These authors contributed equally: Hanna Pulaski, Stephen A. Harrison. ✉e-mail: katy.wack@pathai.com

Methods

Inclusion and ethics

This study included existing de-identified liver biopsies from three clinical trials (Intercept Pharmaceuticals REGENERATE trial NCT02548351, Bristol Myers Squibb FALCON 1 trial NCT03486899 and FALCON 2 trial NCT03486912 and Novo Nordisk semaglutide trial NCT02970942). All patients provided written consent for the original trial; research in this study was granted expedited approval by the WCG Institutional Review Board (IRB00000533).

Datasets and study oversight

The analysis used existing de-identified glass slides and WSIs derived from liver biopsies procured during four MASH clinical trials, including three phase 2 trials and one phase 3 trial (screen failures and enrolled population from the Intercept Pharmaceuticals REGENERATE trial, enrolled population from the Bristol Myers Squibb FALCON 1 trial and FALCON 2 trial and enrolled population from the Novo Nordisk semaglutide trial). These data encompassed a broad spectrum of disease manifestations, captured both screened and enrolled participants and mirrored the variances observed in the MASH clinical trial population. Demographic information was not available for this study; however, all trials enrolled a balanced cohort with respect to sex and gender. Sample collection varied, encompassing historical and study biopsies, with staining procedures executed across multiple sites.

AIM-MASH development

AIM-MASH was trained using 103,579 pathologist-provided annotations of 6,235 H&E and 6,223 Masson's trichrome WSIs from six completed phase 2b and phase 3 MASH clinical trials. For every WSI, AIM-MASH employs a sequential approach in which convolutional neural networks produce tissue overlays containing colorized predictions of segmentation, signifying various histologic features. Additionally, slide-level quantifications of the proportionate area of each feature are generated. Simultaneously, graph neural networks predict an ordinal MASH CRN grade or stage for each histologic feature. The development of AIM-MASH is further described by Iyer et al.[21].

Overlay validation analyses

To assess the accuracy of the heatmap overlays generated by the AIM-MASH model to enable efficient review of key histologic features considered by the algorithm, up to 160 500 × 500-µm-sized frames for each feature (steatosis, lobular inflammation, hepatocellular ballooning, fibrosis, H&E artifact and trichrome artifact) were selected to represent a wide range of each histology and commonly encountered artifacts (for example, tissue folds, stain pooling, scanning blur). Only usable tissue is considered in predicting scores. These overlays are intended to facilitate the pathologist's review in the AIM-MASH scoring workflow and, therefore, were designed with preference for sensitivity. The enrolling pathologist estimated the amount of each feature in each frame on images with no overlays. Three board-certified expert hepatopathologists were provided with the enrolled frames from both H&E and trichrome slides. The pathologists were asked specific questions for each frame to determine to what extent the overlay may or may not be underestimating or overestimating a given feature, defined as TP and FP success rates. Overlay performance was considered acceptable if the TP success rate and the FP success rate were greater than or equal to 85%.

Frames from 222 WSIs were enrolled. Overall, 312 unique H&E frames and 249 trichrome frames were enrolled from three clinical trials (both baseline and follow-up time points from placebo and treatment groups).

Repeatability and reproducibility analyses

For the assessment of AIM-MASH's reproducibility, we incorporated glass slides from two completed phase 2 trials (one non-cirrhotic and one cirrhotic) and a phase 3 MASH trial. To gauge interday repeatability, 150 cases, each comprising an H&E and a trichrome slide, were repeatedly scanned using the same Leica Aperio AT2 scanner at ×40 magnification across three nonsequential days. For intersite reproducibility assessment, identical cases were singularly scanned at three distinct laboratories by different operators using separate AT2 scanners. Reproducibility and repeatability were deemed acceptable when mean pairwise agreement rates consistently matched or surpassed 85%. No pathologist review of AIM-MASH scores was incorporated in repeatability and reproducibility studies.

Establishment of ground truth

GT, defined as the presumed accurate diagnosis, was determined for each case by one of two unique panels of hepatic pathologists. Each panel consisted of two main reader pathologists and an auxiliary tiebreaker pathologist (the tiebreaker was the same between the two panels). Discrepancies in scoring among the primary readers prompted the intervention of the tiebreaker pathologist, who was blinded to initial assessments. When the tiebreaker's scoring diverged from both primary readers, a joint panel call or consensus panel was convened for consensus, with the tiebreaker's score being decisive in rare cases of continued disagreement. Overall, five distinct pathologists contributed to establishing the GT (Fig. 1b). The results from all cases were pooled in the final analysis.

Analytical validation protocol

For analytical validation, 1,481 cases extracted from two finalized phase 2 trials and select cases from a phase 3 trial, representing three different drug candidates with unique mechanisms of action (semaglutide, pegbelfermin, resmetirom), were evaluated in comparison to GT and IMR. Cases from the phase 3 trial were selected to match the original trial enrolled population (baseline and follow-up time points) and included screen failures. Each case underwent scanning via a Leica Aperio AT2 scanner at ×40 magnification. Notably, this phase excluded pathologist review of resultant scores.

Clinical validation protocol

The same cohort of 1,481 cases incorporated in the analytical validation phase was used for clinical validation. This phase aimed to ascertain AIM-MASH's capability to bolster pathologists' accuracy in MASH diagnosis in a therapeutic trial context. The AI-assisted workflow integrated pathologist review of sample quality, staining, scanning adequacy, assessment of any additional findings and subsequent AIM-MASH scoring. Although pathologists could record minor disagreements with AIM-MASH scores, only major discrepancies (two-point or greater difference) permitted score alterations (Fig. 1a) to prevent introduction of interpathologist variability.

Panel comparison

The same cohort of 1,481 cases used in analytical and clinical validation was used to determine the accuracy of AI-assisted reads against two panels of readers. Panel 1 was GT using median scores from the GT readers (median GT) instead of panel consensus calls (Fig. 1c). Panel 2 was the median derived from a minimum of three IMRs (median IMR). Results using these comparisons are described in Fig. 4.

Statistical analysis

Both analytical and clinical validation phases were designed to initially assess AIM-MASH's non-inferiority to manual scoring. Upon confirmation of non-inferiority, its accuracy was further assessed for superiority. Non-inferiority was established when the difference between AIM-MASH Cicchetti–Allison kappa with the GT exceeded a non-inferiority margin of −0.1 compared to the IMR WK with the GT for each MASH component (bootstrap percentile $P < 0.025$). Linearly WK was used, as pairwise comparisons are used to determine the level of agreement and, using this metric, agreement between raters adjusting for the agreement that might occur by chance could be computed. The linear weights, in this case, penalize disagreement due to distant scores (for example, 3 versus 1) more than that between closer ordinal scores (for example, 2 versus 1). This non-inferiority analysis was performed for the algorithm only and AI-assisted results depicted in Fig. 3a,b and for the Fig. 4 results, which compared AI-assisted performance to two different panels. For repeatability and reproducibility, bootstrap percentile *P* values were computed to test the hypothesis that the mean agreement rate, for each MASH component, is greater than 0.85. Additionally, for the post hoc comparisons, where *P* values were not computed, 95% CIs of the point estimates were compared to establish difference.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The histopathology data collected for this study are maintained by PathAI to preserve patient confidentiality and the proprietary image analysis. Access to histopathology features will be granted to academic investigators without relevant conflicts of interest for noncommercial use who agree to not distribute the data. Access requests can be made to A.H.B. (andy.beck@pathai.com). Any additional information required to reanalyze the data reported in this paper relating directly to the clinical datasets (REGENERATE, FALCON 1, FALCON 2 and semaglutide datasets) will be considered at the discretion of the source institute for the clinical trial in question. Requests will be considered from academic investigators without relevant conflicts of interest for noncommercial use who agree to not distribute the data. Data requests should be sent to A.H.B. (andy.beck@pathai.com). PathAI will respond to these requests within 1 month of receipt.

Code availability

Not all original code can be made publicly available. The codes for cell and tissue type model training, inference and feature extractions are not disclosed. To safeguard PathAI's intellectual property, access requests for such code will not be considered. An application for a US patent for the algorithm discussed here has been submitted (WO2022/165433). The source code for all downstream data analyses and figure generation in this work is publicly available and can be downloaded from GitHub at https://github.com/Path-AI/NASH_DDT_Manuscript.

Acknowledgements

We thank all patients and family members that participated in the clinical trials. We also thank all pathologists who participated in the study. We acknowledge the memory of Stephen Harrison, a giant in the field of clinical hepatology who contributed to the advancement of our work and the entire field for the benefit of so many patients in need in countless ways. This study was funded by PathAI, Gilead Sciences, Novo Nordisk, Bristol Myers Squibb and Intercept Pharmaceuticals, and all funders were involved in conceptualization, methodology and preparation of the manuscript.

Author contributions

Conceptualization: H.P., S.A.H., A.J.S., A.S.-M., R.E., C.E.T., R.P.M., C.C., R.P.M., A.-S.S., A.M., V.B., G.M.S., Q.M.A., R.L., V.R., M.C.M., N.P.A., A.H.B., K.E.W. Methodology: H.P., S.A.H., S.S.M., A.J.S., A.S.-M., R.E., C.E.T., R.P.M., C.C., R.P.M., A.-S.S., A.M., V.B., G.M.S., Q.M.A., R.L., V.R., M.C.M., N.P.A., A.H.B., K.E.W. Investigation: H.P., S.S.M., M.C.V., H.H., S.P.M.C., S.M.H., A.S.-M., R.E., J.G., M.R., N.P., C.E.T., K.E.W. Visualization: N.P., H.P. Funding acquisition: N/A. Project administration: H.P., M.C.V., K.E.W. Supervision: M.C.M., N.P.A., A.H.B., K.E.W. Writing (original draft): H.P., S.A.H., S.S.M., A.J.S., M.C.V., L.C.M., H.H., S.P.M.C., S.M.H., A.S.-M., R.E., J.G., M.R., M.R., C.E.T., R.P.M., C.C., S.D.P., A.-S.S., A.M., V.B., G.M.S., Q.M.A., R.L., V.R., M.C.M., N.P.A., A.H.B., K.E.W. Writing (review and editing): H.P., S.A.H., S.S.M., A.J.S., M.C.V., L.C.M., H.H., S.P.M.C., S.M.H., A.S.-M., R.E., J.G., M.R., M.R., C.E.T., R.P.M., C.C., S.D.P., A.-S.S., A.M., V.B., G.M.S., Q.M.A., R.L., V.R., M.C.M., N.P.A., A.H.B., K.E.W.

Competing interests

H.P., H.H., A.S.-M., R.E., N.P., A.H.B. and N.P.A. are full-time, salaried employees of PathAI. K.E.W. was a full-time employee of PathAI during all phases of the study and is now a paid consultant of PathAI. S.A.H. is a paid consultant for Akero Therapeutics, Aligos Therapeutics, Altimmune, Boehringer Ingelheim, Bluejay Therapeutics, Echosens North America, Galecto, Gilead Sciences, GlaxoSmithKline, Hepion Pharmaceuticals, Hepta Bio, HistoIndex, Kriya Therapeutics, Madrigal Pharmaceuticals, Medpace, MGGM Therapeutics, NeuroBo Pharmaceuticals, Northsea Therapeutics, Novo Nordisk, Pfizer, Sagimet Biosciences, Terns and Viking Therapeutics and a shareholder of Akero, Cirius Therapeutics, Galectin Therapeutics, HistoIndex and Northsea Therapeutics. S.S.M., M.C.V., L.C.M., S.P.M.C., S.H.M., C.E.T. and M.C.M. were PathAI employees at the time of study conduct. J.G. and M.R. are paid contractors of PathAI. R.P.M. and G.M.S. are full-time, salaried employees of OrsoBio. C.C. is a full-time, salaried employee of Inipharm. S.D.P. is a full-time salaried employee of Gilead Sciences. A.-S.S. is a full-time, salaried employee of Novo Nordisk. A.M. was a paid consultant for Bristol Myers Squibb. V.B. is a full-time, salaried employee of Bristol Myers Squibb. A.J.S. has stock options in Genfit, Akarna, Tiziana, Indalo, Durect Inversago and Galmed; is a consultant to AstraZeneca, Nimbus, Takeda, Janssen, Gilead, Terns, Merck, Boehringer Ingelheim, Bristol Myers Squibb, Lilly, Novartis, Novo Nordisk, Pfizer and Genfit; and has been an unpaid consultant to Intercept, Echosens, Immuron, Galectin and Affimune Prosciento. His institution has received grant support from Gilead, Bristol Myers Squibb, Intercept, Merck, AstraZeneca and Novartis. He receives royalties from Elsevier and UptoDate. Q.M.A. is a coordinator of the EU IMI-2 LITMUS consortium, which is funded by the EU Horizon 2020 program and the EFPIA. This multistakeholder consortium includes industry partners. He has research grant funding from AstraZeneca, Boehringer Ingelheim and Intercept. He is a consultant on behalf of Newcastle University to Alimentiv, Akero, AstraZeneca, 89bio, Boehringer Ingelheim, Bristol Myers Squibb, Galmed, Genfit, Genentech, Gilead, GlaxoSmithKline, HistoIndex, Intercept, Inventiva, QVIA, Janssen, Madrigal, Merck, NGM Bio, Novartis, Novo Nordisk, PathAI, Pfizer, PharmaNest, Prosciento, Roche and Terns. He is a speaker for Novo Nordisk, Madrigal and Springer Healthcare and receives royalties from Elsevier. R.L. is a consultant to Aardvark Therapeutics, Altimmune, Alnylam–Regeneron, Amgen, Arrowhead Pharmaceuticals, AstraZeneca, Bluejay Therapeutics, Bristol Myers Squibb, Eli Lilly, Galmed, Gilead, Inipharma, Intercept, Inventiva, Ionis, Janssen, Madrigal, NGM Biopharmaceuticals, Novartis, Novo Nordisk, Merck, Pfizer, Sagimet, Theratechnologies, 89bio, Terns Pharmaceuticals and Viking Therapeutics. He is a cofounder of LipoNexus. V.R. is a paid consultant for Novo Nordisk, Northsea Madrigal, Enyo, Poxel, Bristol Myers Squibb, Intercept, NGM Bio and Sagimet.

Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41591-024-03301-2.

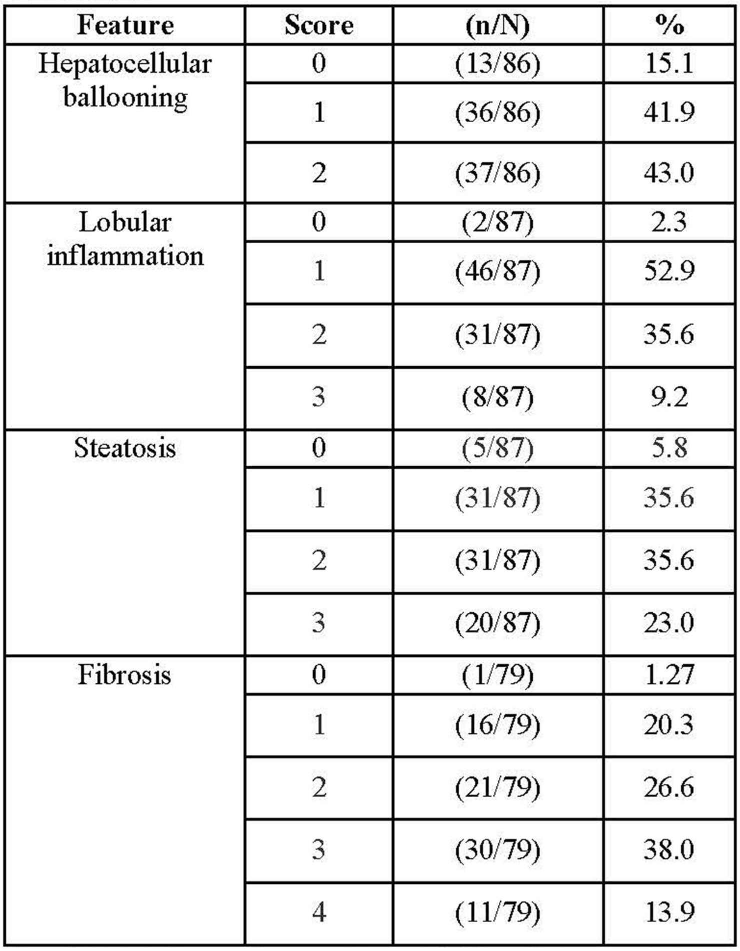
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41591-024-03301-2.

Correspondence and requests for materials should be addressed to Katy E. Wack.

Peer review information *Nature Medicine* thanks Lee Cooper, Mattias Ekstedt and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Lorenzo Righetto, in collaboration with the *Nature Medicine* team.

Reprints and permissions information is available at www.nature.com/reprints.

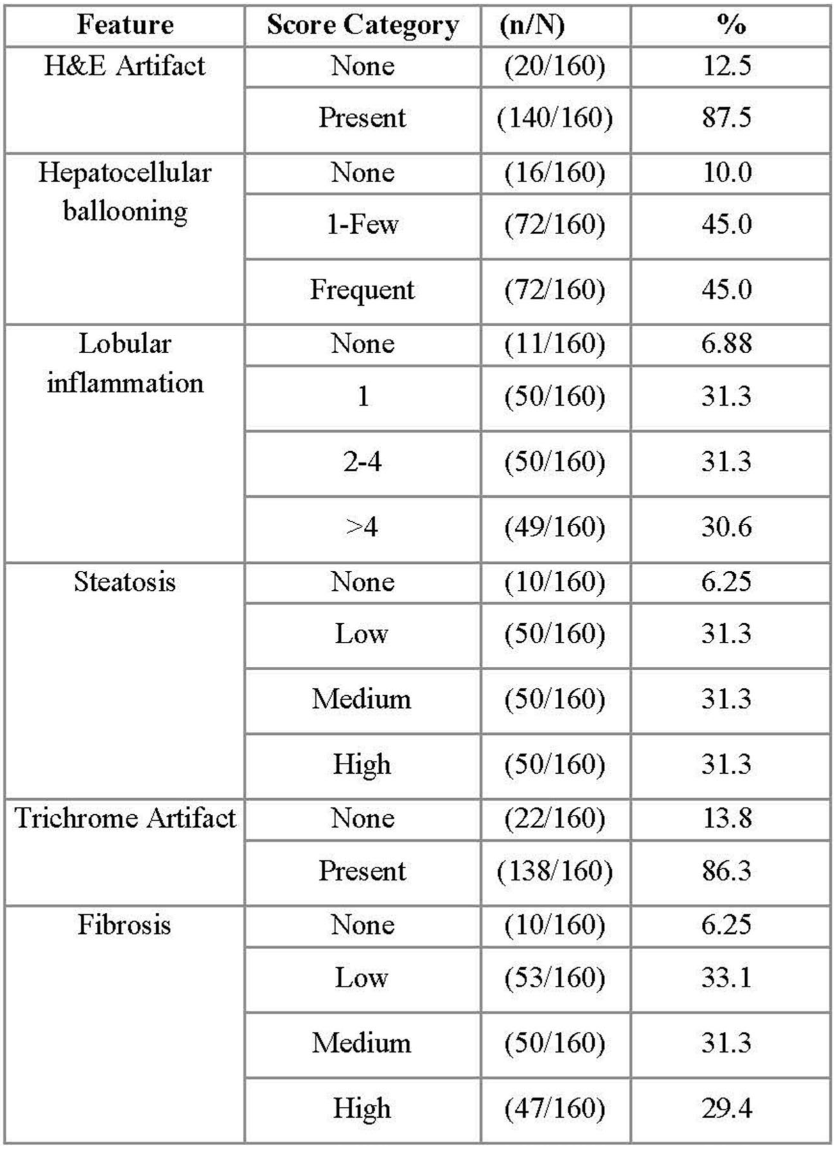
Extended Data Table 1 | Frame Distribution based on Slide Level Score for Overlay Validation



Feature	Score	(n/N)	%
Hepatocellular ballooning	0	(13/86)	15.1
	1	(36/86)	41.9
	2	(37/86)	43.0
Lobular inflammation	0	(2/87)	2.3
	1	(46/87)	52.9
	2	(31/87)	35.6
	3	(8/87)	9.2
Steatosis	0	(5/87)	5.8
	1	(31/87)	35.6
	2	(31/87)	35.6
	3	(20/87)	23.0
Fibrosis	0	(1/79)	1.27
	1	(16/79)	20.3
	2	(21/79)	26.6
	3	(30/79)	38.0
	4	(11/79)	13.9

Multiple frames from within each WSI could be sampled. n=number of WSIs with a particular score within a histologic category from which frames were sampled. N=total number of WSIs for the particular histologic category from which frames were sampled.

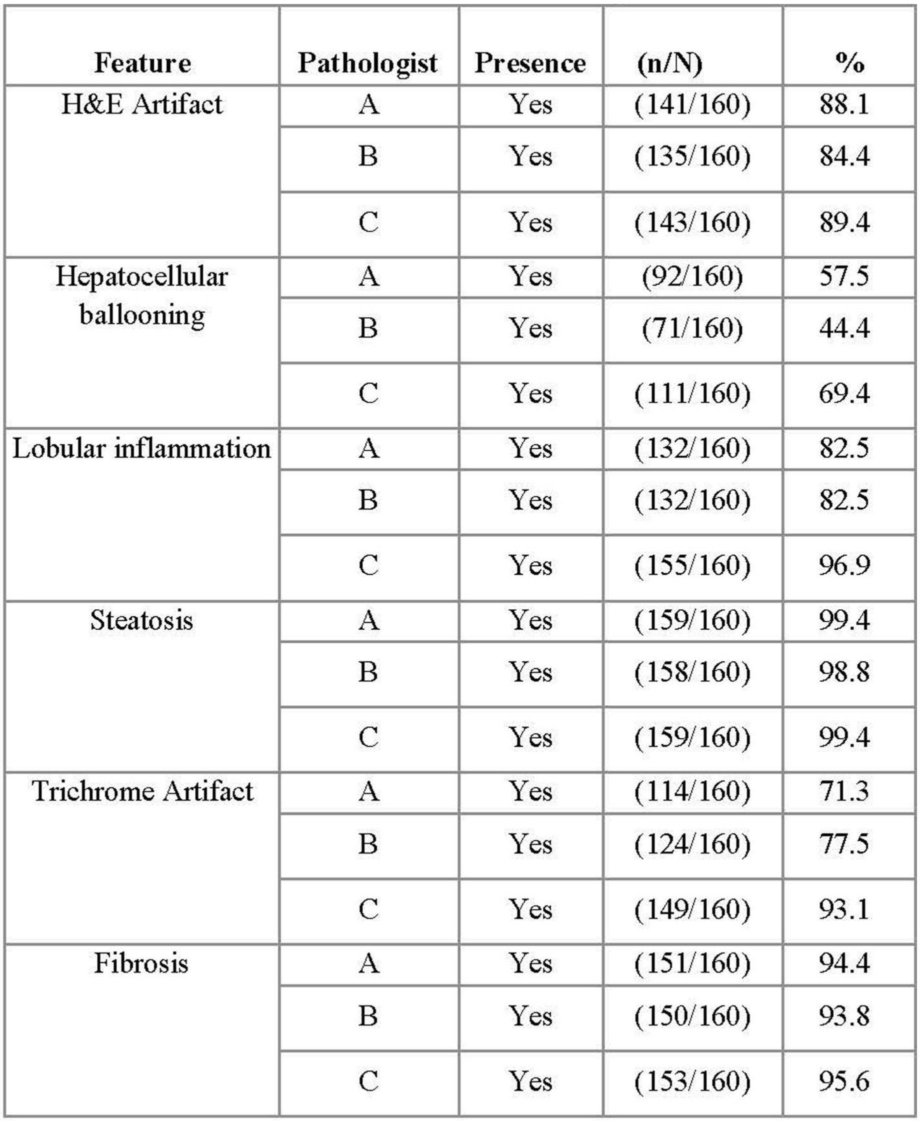
Extended Data Table 2 | Frame Distribution based on Frames Level Score for Overlay Validation



Feature	Score Category	(n/N)	%
H&E Artifact	None	(20/160)	12.5
	Present	(140/160)	87.5
Hepatocellular ballooning	None	(16/160)	10.0
	1-Few	(72/160)	45.0
	Frequent	(72/160)	45.0
Lobular inflammation	None	(11/160)	6.88
	1	(50/160)	31.3
	2-4	(50/160)	31.3
	>4	(49/160)	30.6
Steatosis	None	(10/160)	6.25
	Low	(50/160)	31.3
	Medium	(50/160)	31.3
	High	(50/160)	31.3
Trichrome Artifact	None	(22/160)	13.8
	Present	(138/160)	86.3
Fibrosis	None	(10/160)	6.25
	Low	(53/160)	33.1
	Medium	(50/160)	31.3
	High	(47/160)	29.4

n=number of frames with the indicated score category. N=total number of frames per indicated feature.

Extended Data Table 3 | Presence of Feature per Pathologist for Overlay Validation



Feature	Pathologist	Presence	(n/N)	%
H&E Artifact	A	Yes	(141/160)	88.1
	B	Yes	(135/160)	84.4
	C	Yes	(143/160)	89.4
Hepatocellular ballooning	A	Yes	(92/160)	57.5
	B	Yes	(71/160)	44.4
	C	Yes	(111/160)	69.4
Lobular inflammation	A	Yes	(132/160)	82.5
	B	Yes	(132/160)	82.5
	C	Yes	(155/160)	96.9
Steatosis	A	Yes	(159/160)	99.4
	B	Yes	(158/160)	98.8
	C	Yes	(159/160)	99.4
Trichrome Artifact	A	Yes	(114/160)	71.3
	B	Yes	(124/160)	77.5
	C	Yes	(149/160)	93.1
Fibrosis	A	Yes	(151/160)	94.4
	B	Yes	(150/160)	93.8
	C	Yes	(153/160)	95.6

n=number of frames with feature present per pathologist. N=total number of frames per feature.

nature portfolio

Corresponding author(s): Katy Wack

Last updated by author(s): August 28, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
□	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
□	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection: AI-derived models, input substances, and objectives for application are detailed in the manuscript.

Data analysis: Not all original code can be made publicly available. The code for cell- and tissue-type model training, inference, and feature extractions are not disclosed. To safeguard PathAI's intellectual property, access requests for such code will not be considered. An application for a United States patent for the algorithm discussed herein has been submitted (WO2022/165433). The source code for all downstream data analyses and figure generation in this work are publicly available and can be downloaded from GitHub: https://github.com/Path-AI/NASH_DDT_Manuscript. Software utitlized in collection and analysis: OpenClinica (Electronic Data Capture system), v13.2.2, v13.3.1, PathAI AISight Research and AISight Clinical Trial Platforms v2.0.0, v2.1.1, 3.0, and 3.1 were utilized for whole slide image viewing and algorithm review data capture. AIM-MASH algorithm v1.1.0 was utilized for this study. All software updates were minor and follow strict change control procedures to ensure no updates effect trial data capture, analysis, and integrity in general. Finally Python v3.7.5 was the primary analysis package utilized for statistical analyses.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The histopathology data collected for this study is maintained by PathAI to preserve patient confidentiality and the proprietary image analysis. Access to histopathology features will be granted to academic investigators without relevant conflicts of interest for non-commercial use who agree not to distribute the data. Access requests can be made to Andrew Beck (andy.beck@pathai.com). Any additional information required to reanalyze the data reported in this paper relating directly to the clinical datasets (REGENERATE, FALCON 1, FALCON 2, and Semaglutide datasets) will be considered at the discretion of the source institute for the clinical trial in question. Requests will be considered from academic investigators without relevant conflicts of interest for non-commercial use who agree not to distribute the data. Data requests should be sent to Andrew Beck (andy.beck@pathai.com). PathAI will respond to these requests within one month of receipt.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	No sex and gender data was available for the samples used. These samples were all completely de-identified samples from completed clinical trials. The only information available for the datasets were sample collection time point (baseline vs post-baseline)
Population characteristics	No data was available for the samples for race, ethnicity or other socially relevant groupings. These samples were all de-identified samples from completed clinical trials
Recruitment	Not applicable. All samples were from completed clinical trials.
Ethics oversight	The research was granted expedited approval by the WCG Institutional Review Board (IRB00000533).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For analytical validation: Six-hundred (600) patients were selected based on to ensure 90% power across component scores for accuracy. To generate these sample sizes, the following methods were used. Given a range of scores for each NASH component, inter-pathologist Kappa based on literature, and non-inferiority margin of 0.1, both the upper bound of inter-pathologist Kappa (Target) at 90% power and lower bound (LB) of Kappa between AIM-NASH model and consensus evaluated during internal testing studies at alpha of 0.025 were estimated based on the parametric model. Simulations were run to find the smallest N for which the below test passed: LB > Target - 0.1 Based on the CRN literature (Kleiner et al. 2005), different inter-pathologist Kappas were expected for different components. Sample size in is computed based on the simulation described above and interpathologist Kappa in the second column of the table. Thus, a sample size of 600 was selected since this provides the most conservative estimate for the component with the most variable interpathologist Kappa - hepatocytic ballooning with a Kappa of 0.5. For reproducibility and repeatability: Assuming a mean agreement rate of 92% for reproducibility/repeatability based on internal pilot data, and a target of 85% at one-sided alpha of 0.025 based on Wilson score confidence interval, a sample size of 120 will be needed to achieve a power of at least 95%. A sample size of 150 ensures adequate sample size to account for any slides which may be broken in shipment or handling across the three reproducibility sites. Clinical validation: Cases from 3 completed clinical trials were included. All available cases from 2 phase 2 studies were included to provide evidence of accuracy across a range of trial phases, drug candidates, and central labs used for staining and scanning. A subset of a phase 3 trial was also included and a sample size of 600 cases were selected for the phase 3 trial based on power calculations to ensure 90% power across component scores. Based on the sample sizes available from these studies, the overall enrollment goal was 1424 cases. Overlay validation: Assuming a success criteria rate of 95% and a target of exceeding 85% at one-sided alpha=.025 level for a single rater. Wilson score confidence intervals then gives N (for each overlay) of 115 or 138 at 90% or 95% power, respectively. In order to ensure adequate power, and account for the possibility of missing features, the N for each feature will be at least 160 frames.
Data exclusions	Missing data could occur if a glass slide broke or pathologists deemed the slides to be not evaluable due to various reasons (wrong stain, wrong slide, wrong scanner, broken slide, poor scan quality, poor stain quality, sample inadequacy, poor sample evaluability). Analysis was

performed on complete case basis separately for each of the 4 NASH components. Cases with missing values for a particular NASH component from ground truth (GT), individual manual read (IMR), or AI-assisted were excluded from analysis of that NASH component.
For analytical validation - Each case for repeatability and reproducibility has 3 AIM-NASH scores. In cases where
AIM-NASH was not able to run on a slide, the slide was removed from analysis. Out of the 607 enrolled slides, less than 4% of the slides had missing final GT score due to various reasons (such as sample, stain or scan adequacy). There were no slides where all IMR pathologists were unable to score the slide for all components. One slide was deemed inadequate for AIM-NASH.
For clinical validation - Out of the 1501 enrolled slides, less than 4% of the slides had missing final GT score due to various reasons (such as sample, stain or scan adequacy); most being for fibrosis (3.2%) and least being for steatosis (1.33%). Similarly, less than 1% of the slides had a missing score from all IMR pathologists reviewing the slide for all components. Additionally, less than 4% of the slides had a missing score from the AI-assisted workflow due to the pathologists unable to score the slide. There were 7 slides where AIM-NASH was not able to provide a score due to blurry images.
There were no missing data for overlay validation.
All 3 protocols had predefined inclusion and exclusion criteria.

Replication

All data analysis of the collected data was performed by two biostatisticians independently who matched their results in the end.

Randomization

NA, no randomization was performed, as no intervention was involved. The study team was also blinded to the treatment groups of the patients from the original trials

Blinding

NA. All study staff was blinded to data during the study, except for clinical data managers and unblinded clinical scientists for the purpose of query identification and resolution.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
□	☒ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Clinical data

Policy information about clinical studies
All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration

NA

Study protocol

NA, submitted with the manuscript

Data collection

Prospective reads of retrospective clinical trial samples was performed in this study. Data collection was done by pathologists who reviewed the histology slides either manually or with AI-assistance and entered their scores into a electronic data capture system.

Outcomes

All primary and secondary endpoints were defined in study protocols. Additional exploratory analysis was performed.