EMA Novel Methodology Qualification Briefing Document

SUBMISSION TITLE: ***AI-Based Histologic Measurement of NASH (AIM-NASH)***

Administrative Information

REQUESTING ORGANIZATION:
PathAI
https://www.pathai.com/

Organization Address:
1325 Boylston St, Suite 10000, Boston, MA 02215, USA
Phone: +1 617-500-8457
Email: N/A

Primary Contact:
Name: Katy Wack
Address: 1325 Boylston St, Suite 10000, Boston, MA 02215, USA
Phone: +1 412-728-1217
Email: katy.wack@pathai.com

Alternative Contact:
Name: Nick Anderson
Address: 1325 Boylston St, Suite 10000, Boston, MA 02215, USA
Phone: +1 571-242-6589
Email: nick.anderson@pathai.com

SUBMISSION DATE: August 24, 2023

1 Table of Contents

Table of Contents

1.1 Acronyms and Definitions

Term	Definition
ABMS	American Board of Medical Specialties
AI	Artificial intelligence
AI-assisted	AIM-NASH workflow, where the pathologist reviews the AIM-NASH scores
AIM-NASH	Artificial Intelligence-based Measurement of Nonalcoholic Steatohepatitis
AV	Analytical Validation
AWS	Amazon Web Services
CAP	College of American Pathologists
CI	Confidence Interval
CLIA	Clinical Laboratory Improvement Amendments
CNN	Convolutional Neural Networks
Contributor Network	A network of over 400 pathologists contracted to provide a wide variety of

	pathology services to PathAI. These pathologists come from diverse backgrounds (academic medical centers, private practices etc.) with a variety of experience (from newly out of fellowship to experts in their fields).
CRF	Case Report Form
CRN	Clinical Research Network
CRO	Contract Research Organization
CTS platform	The "Clinical Trial Services Platform" is a research use only (RUO) cloud-based software as a service (SaaS) that enables PathAI and partners to conduct their clinical trials either with or without the use of PathAI clinical trial algorithms. This platform was previously called "Clinical Trials Portal." The CTS platform has recently been re-named to AISight Clinical Trials platform and is referred to as such herein.
CV	Clinical Validation
DZI	Deep zoom image
EC2	Elastic Compute Cloud
eDC	Electronic Data Capture
EKS	Elastic Kubernetes Service
EMA	European Medicines Agency
eQMS	Electronic Quality Management System
FDA	Food and Drug Administration
FN	False Negative
FFPE	Formalin-fixed, paraffin-embedded
FP	False Positive

GDPR	General Data Protection Regulation
GNN	Graph Neural Networks
GT	Ground Truth
H&E	Hematoxylin and eosin
HBV	Hepatitis B Infection
IA	Intermediate Analysis
IAM	Identity and Access Management
IAV	Integrated Analytical Verification
ICH GCP	International Conference on Harmonisation Good Clinical Practice
IFU	Instructions for Use
IMR	Independent Manual Read
ISMS	Information Security Management System
IRB	Institutional Review Board
LB	Lower Bound
ML	Machine learning
NAFLD	Nonalcoholic fatty liver disease
NAS	NAFLD Activity Score
NASH	Nonalcoholic steatohepatitis
NASH scoring system (CRN Fibrosis Stage and NAS Score)	Histologic scoring system developed by NASH CRN (1). This scoring system comprises of NAFLD Activity Score (NAS) and fibrosis stage. The NAS comprises of steatosis (0-3), lobular inflammation (0-3) and hepatocyte ballooning (0-2) and is scored using an H&E-stained slide. Fibrosis stage ranges

	from 0 to 4 and is provided using a trichrome-stained slide.
NI	Non-Inferiority
PHI	Protected Health Information
PI	Principal Investigator
PSC	Primary Sclerosing Cholangitis
PV	Platform Validation
QC	Quality Control
Qualification	The act of proving and documenting that equipment or ancillary systems are properly installed, work correctly, and comply with specified requirements. The process is used to demonstrate the ability to fulfill pre-specified requirements for a task or a process. (ICH7 Good Manufacturing Practice Guidance)
RDS	Relational Database Service
RUO	Research Use Only
SaaS	Software as a Service
SAV	Standalone Analytical Verification
Slides platform	The Slides platform is a research use only (RUO) cloud-based software that enables PathAI to develop and test algorithms and in rare cases, partners to utilize the platform in retrospective clinical trials. Platform configurability allows for maximum flexibility in leveraging digital pathology to improve outcomes in translational and clinical research. The Slides platform has recently been re-named to AISight

	Translational platform and is referred to as such herein.
SOC	Standard-of-Care
SOP	Standard Operating Procedure
SQS	Simple Queue Service
TP	True Positive
UI	User Interface
WK	Weighted Kappa using linear weights (Cicchetti-Allison weighted kappa)
WSI	Whole Slide Imaging

Tables and Figures

2 Executive Summary

2.1 Objectives

PathAI has recently shown that a machine learning (ML) approach, Artificial Intelligence-Based Histologic Measurement of Nonalcoholic Steatohepatitis (AIM-NASH), can be used to support pathologic interpretation of NASH liver biopsies by accurately and efficiently quantifying NASH histologic features. The current method by which pathologists apply NASH histologic scoring systems is subjective and prone to variability, whereas the ML-assisted approach is more accurate and reproducible. Therefore, AIM-NASH has the potential to be used to aid clinical trial enrollment and histologic endpoint assessment, providing value for accelerated and traditional NASH drug approval pathways.

The purpose of this document is to apply for EMA qualification of this novel AIM-NASH methodology for use as an aid to pathologists in NASH clinical trials.

2.2 The need and impact for AIM-NASH as a drug development tool

Background

Nonalcoholic fatty liver disease (NAFLD) is rising in prevalence globally, with an estimated 25% of the world's population affected (1). Due to this increased burden of disease, liver decompensation due to progression of NAFLD to nonalcoholic steatohepatitis (NASH) is expected to become the leading cause of liver transplant (2). There are currently no approved therapies for NASH and there is a large unmet need for clinical intervention in this patient population. While this has been an active drug development area, progress has been stymied by numerous failed trials with borderline results. In this context, it is important to remember that there are true human costs associated with the lack of progress in this critical area and therapeutic options are desperately needed. Many patients go undiagnosed until late-stage disease and often struggle to establish meaningful lifestyle changes while waiting for transplant and can decompensate quickly with no treatment options.

A major obstacle in the development of effective NASH therapies is the lack of a reliable histologic scoring method to both identify patients for trials, and more importantly, to assess the effectiveness of experimental interventions in a reasonable period. The current gold standard to establish the diagnosis of NASH is histologic analysis of the liver biopsy. The diagnosis of NASH is established by the presence of a characteristic histologic pattern of elevated steatosis, inflammation, and hepatocellular ballooning on liver biopsies in the absence of significant alcohol intake, with patterns and extent of fibrosis being assessed for staging. Steatohepatitis, as opposed to steatosis, is a progressive disease which ultimately leads to cirrhosis. In 1999, a semi-quantitative grading and staging system to describe and unify the approach of pathologists to the histopathologic lesions of NASH was proposed by Brunt et al (3). A semi-quantitative activity grade was assigned by a combination of parameters including steatosis, lobular and portal inflammation, and hepatocyte ballooning. Fibrosis staging was based on fibrosis patterns of NASH in adults reflecting progression of fibrosis to end stage cirrhosis.

Change in histologic features observable on liver biopsy has been acknowledged by FDA as reasonably likely to predict clinical benefit (4). Calculation of histologic change can therefore be used as surrogate endpoint in NASH trials, given that clinically relevant outcomes (evidence of decompensation, liver transplant, and mortality) may take years to develop. While liver biopsy is invasive and may not yield a representative section of liver tissue, it is still the best means to monitor patients with NASH as histologic changes have been shown to be highly predictive of long-term outcomes (5). The NAFLD activity score (NAS) established by the NASH Clinical Research Network (CRN) measures disease activity through scoring of steatosis, hepatocellular ballooning and lobular inflammation (Table 1 and Table 2) (6). Improvement in NAS score has been associated with an improvement in fibrosis (6). The fibrosis scale, also developed by the CRN, assesses worsening fibrosis or progression toward cirrhosis and conversely improvement in fibrosis, and is a strong predictor of long-term outcomes (7).

Table 1: NAFLD Activity Score Components Developed by NASH Clinical Research Network (CRN). Adapted from (8)

Item	Definition	Score
Steatosis	< 5%	0
	5%-33%	1
	> 33%-66%	2
	> 66%	3
Lobular inflammation	No foci	0
	< 2 foci per 200 x field	1
	2-4 foci per 200 x field	2
	> 4 foci per 200 x field	3
Hepatocellular ballooning	None	0
	Few balloon cells	1
	Many cells/ prominent ballooning	2

Table 2: CRN-developed Fibrosis Scoring System. Adapted from (8)

Fibrosis Stage	
0	None
1A	Mild, zone 3 perisinusoidal
1B	Moderate, zone 3 perisinusoidal
1C	Periportal sinusoidal fibrosis without accompanying zone 3 fibrosis
2	Zone 3 perisinusoidal and portal/periportal
3	Bridging fibrosis
4	Cirrhosis

Various sets of semi-quantitative pathologic criteria have been proposed for scoring NASH (9,10). In their guidelines for clinical trials, both Food and Drug Administration (FDA) and European Medicines Agency (EMA) have adopted the NAS and CRN fibrosis measurement schemes to define surrogate endpoints that have a reasonable likelihood to evaluate clinical benefits (11–13).

NAS and CRN Fibrosis scoring systems are the current gold standard measurements for NASH, and existing manual histologic scoring systems have suboptimal reproducibility, even when used by expert hepatopathologists (see

Table 3, Table 4, Table 5) in Impact of Methodology Section (8,14–17). In addition to the risk of inter- and intra-reader variability, NASH trials are subject to systematic bias in the form of temporal bias. Methods used to guard against systematic bias include randomization of slide assignment (mixing baseline and follow-up samples) and assignment of multiple central pathologists. However, the possibility of systematic bias is more difficult to control when grading and staging are required for trial entry.

Misclassification of NAFLD activity and fibrosis staging has major downstream effects on clinical trial results. Measurement error due to inconsistency in manual sample reading can lead to inadequate power due to the misclassification of true responders vs. non-responders during endpoint analyses, which ultimately leads to inaccurate interpretation of trial results. This can lead to the potential failure of an efficacious drug. An evaluation of results from the EMMINENCE study exemplifies this problem.

The EMMINENCE study was a Phase 2 trial evaluating the effects of MSDC-0602K on liver histology which failed to meet the FDA defined liver histology endpoints (18). In a follow-up evaluation, two hepatologists independently read the baseline and follow-up liver biopsies which were compared to the original central pathologist reading. The WKs for inter-rater reliability for each NAS component and fibrosis score ranged from 0.328 to 0.609. The key finding from this study was that 46.3% of subjects enrolled in EMMINENCE were found to be ineligible by at least one of the three hepatologists during this retrospective evaluation. Misclassification of NAFLD activity and fibrosis staging was also found to reduce study power from 90% to 40% in simulations (14). To address these inconsistencies, trials often implement processes involving multiple reviewers or blinded review. This adds additional time to the trial workflow and personnel needed to complete scoring. In addition, this does not address the inconsistency across reads. An unbiased and reproducible method for trial enrollment and endpoint evaluation is needed to yield reliable results from NASH trials.

Over the last five years, artificial intelligence (AI) and ML-based tools have shown enormous potential and progress in medical specialties like ophthalmology, radiology, pathology, oncology, and general medical decision support. 240 AI/ML based, CE marked medical devices and algorithms were identified between 2015 and March 31, 2020. Of these devices, only two devices were classified as risk class III (19).

In summary, currently, there are no approved therapies for NASH. Due to its increasing prevalence, NASH is expected to soon become the leading cause of liver transplant (2,20,21). There is a critical need for a scalable, reproducible, and validated tool in quantitative pathology for the assessment of efficacy in NASH therapeutic development.

Impact of proposed methodology

Existing manual histologic scoring systems for NASH have suboptimal reproducibility, even when employed by expert hepatopathologists such as the NASH Clinical Research Network (CRN) pathology committee (

Table 3). Inter-reader agreement assessed by Kappa statistics is only moderate for lobular inflammation, hepatocellular ballooning, and overall NAFLD activity score (NAS). Inter-reader agreement for fibrosis and steatosis is substantial, but still indicates variability between pathologists. These deficiencies have persisted over time, despite educational and training efforts to improve them (6,8). Suboptimal intra-reader agreement for many of these parameters is also an issue, with variable rates of agreement related to design considerations of re-read and paired read comparisons. (Table 4, Table 5, Table 6).

Table 3: Inter-reader Agreement for NASH Histologic Features from the NASH-CRN

Feature	**Kleiner 2005** (22) **Kappa Statistics**	**Kleiner 2019** (22) **Kappa Statistics (95% CI)**	**Davison 2020** (14) **Kappa Statistics**	**Davison 2020** (14) **Average % agreement**
N	32 Cases	446 Cases	678 Cases	678 Cases
Steatosis	0.79	0.77 (0.69-0.84)	0.609	63.32%
Lobular inflammation	0.45	0.46 (0.34-0.58)	0.328	60.37%
Hepatocellular ballooning	0.56	0.54 (0.44-0.65)	0.517	62.54%
Fibrosis	0.84	0.75 (0.67-0.82)	0.484	50.93%
NAFLD Activity Score (NAS)	-	0.52 (0.44-0.60)	0.495	32.25%
'Gestalt' NASH Diagnosis	0.61	0.66 (0.57-0.75)	0.399** (0.340-0.459)	79.60%

**Kleiner NASH diagnosis is "gestalt"*

***Davison NASH diagnosis is an unweighted Kappa, all others in table are weighted.*

Table 4:Intra-reader Agreement for NASH Relevant Score

Feature	**Kleiner 2005** (8) **Kappa Statistic**
N	*32 Cases*
Steatosis	0.83
Lobular inflammation	0.60
Hepatocellular ballooning	0.66
Fibrosis	0.85

Table 5: Intra-reader Percent Agreement for NASH Histologic Features, qualifying vs re-read in a clinical trial population (14)

Feature	**Pathologist A % Agreement**	**Pathologist A WK (95% CI)**	**Pathologist A % Agreement**	**Pathologist A WK (95% CI)**
N	*389*	*389*	*200*	*200*
Steatosis	72.24%	0.666 (0.609-0.722)	69.50%	0.625 (0.543-0.708)
Lobular inflammation	55.27%	0.227 (0.154-0.300)	56.50%	0.244 (0.142-0.347)
Hepatocellular ballooning	69.92%	0.487 (0.423-0.552)	71.00%	0.497 (0.404-0.590)
Fibrosis	71.98%	0.679 (0.625-0.733)	73.50%	0.720 (0.654-0.787)
NAFLD Activity Score (NAS)	37.02%	0.372 (0.318-0.427)	39.00%	0.372 (0.294-0.449)

Table 6: Intra-reader Agreement for NASH Histologic Features, Individual vs Paired Read (14)

Feature	**Pathologist B (vs paired screening read) % Agreement**	**Pathologist B (vs paired screening and 12 month reads) WK (95% CI)**	**Pathologist B (vs paired screening read) % Agreement**	**Pathologist B (vs paired screening and 12 month reads) WK (95% CI)**
N	*200*	*200*	*400*	*400*
Steatosis	89.50	0.861 (0.804-0.918)	88.00	0.863 (0.825-0.901)
Lobular inflammation	75.50	0.633 (0.52-0.724)	73.75	0.662 (0.604-0.720)
Hepatocellular ballooning	86.00	0.821 (0.758-0.884)	86.25	0.840 (0.799-0.881)
Fibrosis	86.50	0.876 (0.832-0.921)	84.50	0.854 (0.819-0.890)
NAFLD Activity Score (NAS)	61.50	0.718 (0.656-0.779)	58.75	0.758 (0.721-0.794)

Recently, to help to solve the individual reader variability challenges, trials have employed more than one reader per case, with varying consensus workflows. These panels, while likely reducing some of the individual bias, still have limitations, including inter-panel variability, temporal bias, lack of standardization across trials, increased time/complexity and increased costs. Some workflows involve 2 independent pathologists scoring, followed by consensus sessions to resolve any discrepant scores and involving a third pathologist in rare cases where consensus cannot be met. This method results in the need

for many consensus sessions, given the high rate of inter-pathologist variability, with another limitation being that the outcome can be influenced by the pathologist reading style (e.g., both being similar or different in their definition of ballooning). Other consensus workflows employ three independent readers, utilizing modal agreement and only fully discrepant scores are taken to consensus. This method is also impacted by the scoring styles of the chosen readers. This is not surprising since the pairwise Kappas of NASH experts vary widely and are different from those averaged across larger numbers of CRN pathologists (22). This inter-panel variability has been observed and was demonstrated in a pilot study using a subset of biopsies (N=100) from the REGENERATE trial. The panels were composed of three different qualified NASH pathologists each (Table 7), where all pathologists initially provided independent reads. Final scores were determined per each component where at least two pathologists in a panel agreed on a histologic score, or by consensus sessions when all three initial pathologist scores were discrepant. These inter-panel agreements were calculated and compared to the Kappas demonstrated by the CRN and other groups in the literature. While there is a slight improvement in variability using panels for some histologic features, overall variability remains across all features, and especially for hepatocyte ballooning.

Table 7: Inter-Panel Variability Remains in Gold Standard Consensus Panels

Feature	**Shrout-Fleiss WK (using quadratic weights)**			**Cicchetti-Allison WK (using linear weights)**		
	Sanyal 2021 (23) Panel A vs B (N=100)	Kleiner 2019 (22)[a] (N=446)	Kleiner 2005 (8)[a] (N=32)	Sanyal 2021 (23) Panel A vs B (N=100)	Davison 2020 (14)b (N=339)	Newsome 2021(24)[b,c] (N=320)
Fibrosis	0.82	0.75	0.84	0.71	0.48	0.61-0.65
Lobular inflammation	0.60	0.46	0.45	0.46	0.33	0.38-0.39
Hepatocellular ballooning	0.62	0.54	0.56	0.51	0.52	0.41-0.61
Steatosis	0.89	0.77	0.79	0.83	0.61	0.63-0.76

N, number of patients.
Dataset is a subset of biopsy samples from enrolled patients from the Regenerate ph3 trial in a prospectively read, retrospective analysis.
Results from current analysis are based on non-missing values.
Panels A, B, consistent of 3 expert NASH pathologists each. Independent reads were collected from each of the 3 for each component. A score was considered to be final if 2 out of 3 reads agreed. If there was complete discrepancy, the 3 pathologists met as a panel to come to consensus.
[a]Average of pairwise Kappas.
[b]Pairwise Kappas.
[c]Range based on 2 values from baseline and week 72 slides

Misclassification of NAS component grading and fibrosis staging has potentially significant consequences for evaluation of clinical trial results. Measurement error due to variability in manual sample reading can lead to: a) inconsistent enrollment of trial subjects who meet inclusion criteria, b) inadequate study power, c) exaggerated apparent response in placebo arm, and d) misclassification of responders vs. non-responders. Each of these factors may ultimately contribute to inaccurate interpretation of results and potentially denial of efficacious therapeutics.

The use of AIM-NASH mitigates the current risks of manual scoring in NASH trials. The following benefits may be expected:

- Accurate and precise scoring of disease activity, standardized across trials (phase, drug class, disease severity).
- Increased likelihood of trial success with reproducible metrics enabling histologic detection of drug response thereby preventing non-approval of efficacious drugs.
- Decreased likelihood of inaccurate endpoint analysis and invalid study conclusions impacting drug approval decisions.
- Scalable method to accelerate drug development.

2.3 Characteristics of the AIM-NASH tool

Biomarker Name: AI-Based Histologic Measurement of NASH (AIM-NASH)
Type of Biomarker: Histologic based, imaging modality, measurement based on machine learning.
BEST Classification: Monitoring Biomarker

The biologic entities measured by the AIM-NASH biomarker are the histologic features comprising the nonalcoholic fatty liver disease (NAFLD) activity score (NAS - steatosis, lobular inflammation, hepatocellular ballooning) and fibrosis stage, as identified by pathologists on whole slide images (WSIs) of liver biopsy slides. Tissue regions of steatosis, lobular inflammation and hepatocellular ballooning are identified on WSIs of hematoxylin and eosin (H&E) stained slides (Table 1) and graded individually using the NASH CRN scoring system (8). Areas of fibrosis are identified on WSIs of trichrome-stained biopsies and staged using the NASH CRN scoring system (Table 2). Biomarkers are measured using a pipeline of ML algorithms that detect and score key NASH histologic features.

2.4 Context of Use for which a Qualification is requested

A monitoring biomarker as an adjunct that aids the pathologist in assessing NAS score (steatosis, hepatocellular ballooning, lobular inflammation) and fibrosis stage (at baseline and follow-up time points) in liver biopsies in NASH clinical trials.

The AIM-NASH outputs will mirror the current EMA guidelines for NASH evaluation for enrollment in clinical trials, measurement at follow-up time points and histologic endpoint evaluation. AIM-NASH is accessible to users on a web-based platform, integrating into NASH clinical trials without significant impact to workflow or introducing new risk to patients. The pathologist will review the output of AIM-NASH and take an active role in its interpretation by accepting or rejecting each of the NAS components

and fibrosis stage, after confirming sample evaluability and determining the presence of any additional findings. AIM-NASH should be used with slides scanned with the Aperio AT2 scanner.

The biomarker is applicable to screening and at follow-up time points for phase 2 and phase 3 NASH trials. This includes patients with fibrosis stages ranging from 0-4 and NAS <4 and ≥4.

2.5 Sources of data and major findings

2.5.1 Validation of the AISight Clinical Trials and Translational Platforms: Purpose and Methods

The purpose of the digital platform validation studies is to evaluate if NASH reads performed using WSIs scanned with Aperio AT2 RUO (Aperio AT2) scanner at 40x on the AISight Clinical Trials platform and the AISight Translational platform are non-inferior to the NASH reads performed using glass slides. Note in the accepted Final Briefing Document for Qualification Advice delivered by EMA on November 19, 2021, the terms PathAI Clinical Trial Portal and Slides platform were used. These have since been changed to AISight Clinical Trials platform, also previously referred to as Clinical Trials Services (CTS) platform, and AISight Translational platform, respectively. The AISight Translational platform is an additional research viewer that was utilized to collect manual digital reads and ground truth reads in validation studies.

This study first established ground truth utilizing median NASH scoring across 3 ground truth (GT) hepatopathologists. The study evaluated for non-inferiority of digital reads (on each AISight Clinical Trials Platform and Translational Platform) to glass reads in agreement with ground truth. For the purposes of this study, NASH is defined as NAS $\geq$ 4 with a score of $\geq$ 1 for each component: steatosis, lobular inflammation and hepatocellular ballooning and absence of atypical features suggestive of non-NASH liver disease. Digital read accuracy was evaluated for non-inferiority to glass read accuracy by comparison to ground-truth scores with a non-inferiority margin of 0.05.

2.5.1.1 Data Sources

PathAI utilized 320 existing de-identified glass slides (160 cases) inclusive of slides from a third-party vendor and from partners from their completed clinical trials (screen failures from Phase 2B study and enrolled cases from a different Phase 2 study).

2.5.1.2 Major Findings by Platform and Conclusions

2.5.1.2.1 AISight Clinical Trials Platform

The validation of AISight Clinical Trials platform demonstrates that the accuracy of NASH digital reads on the platform for slides scanned on Aperio AT2 (Leica) scanner are non-inferior to reads performed with traditional light microscopy with glass slides. This study demonstrated a significant non-inferior average agreement (across three pathologists) of manual digital reads with GT as compared to glass read agreement with GT (NI margin of 0.05, difference of -0.001, 95% CI of (-0.027,0.026), and $p<0.0001$). Additionally, for each individual pathologist, there was a similar average agreement of manual digital and glass reads with glass GT (pathologist A 0.843 and

0.849, pathologist B 0.633 and 0.605 and pathologist C 0.755 and 0.780). Average intra-reader, inter-modality (glass to digital) WKs for each score component in this study were higher than WKs in published literature (14). The results from this NASH digital platform validation study support the conclusion that digital reads performed on the Clinical Trials Platform are non-inferior to glass reads in reference to glass GT when used by pathologists for NASH assessment in a clinical trial population.

2.5.1.2.2 AISight Translational Platform

The validation of the AISight Translational platform demonstrates that the accuracy of NASH digital reads on the platform for slides scanned on Aperio AT2 scanner are non-inferior to reads performed with traditional light microscopy with glass slides. This study demonstrated a significant non-inferior average agreement (across three pathologists) of manual digital reads with GT as compared to glass read agreement with GT (NI margin of 0.05, difference of -0.004, 95% CI of (-0.045,0.036), and p=0.0110). In addition, for each individual pathologist, there was a similar average agreement of manual digital and glass reads with glass GT (pathologist A 0.759 and 0.744, pathologist B 0.750 and 0.801 and pathologist C 0.859 and 0.821). Overall and per pathologist intra-reader, inter-modality (glass to digital) WKs for each score component in this study were within the range or close to the lower bound of published WKs, with higher NAS intra-reader agreement for all pathologists in this study compared to the published values from (published WK range for steatosis 0.666-0.83, for lobular inflammation 0.227 – 0.60, for hepatocellular ballooning 0.32-0.66, for fibrosis 0.64-0.85, for NAS value 0.372; (8,14,25). The results from this NASH digital platform validation study support the conclusion that manual digital reads performed on the AISight Translational platform are non-inferior to the glass reads in reference to glass GT when used by pathologists for NASH assessment in a clinical trial population.

2.5.1.2.3 Conclusions

The data from the validation of both AISight Clinical Trials and Translational Platform demonstrate glass-to-digital equivalence for biopsy evaluations for NASH clinical trials in a challenging, borderline population. The overall WKs compared to GT were slightly different between platforms, but these values were all within the expected inter-reader range for each histologic component and for total NAS. The variation can likely be attributed to the utilization of different panel readers for each validation study and the known variability between readers, as demonstrated in the literature (8,14,25). Importantly, glass to GT and digital to GT read agreements were equivalent for both platforms.

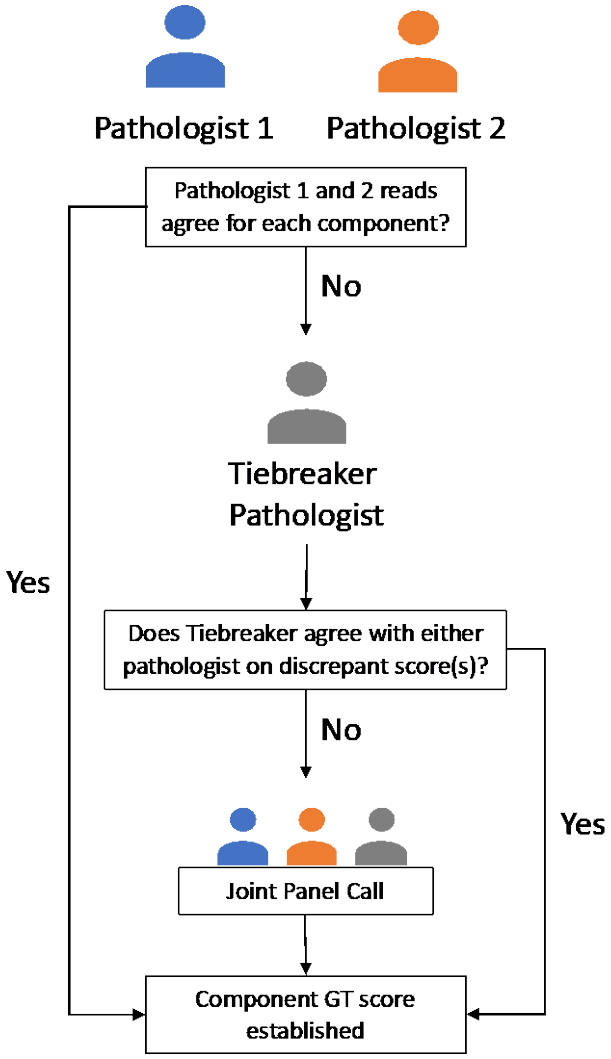
Incorporating digital pathology into clinical trial workflows makes trial management more efficient, allows for multiple reads in parallel, and provides opportunities to utilize the most experienced pathologists for reader panels as geographic location is no longer a factor for selecting pathologists. Additionally, with the ongoing development and validation of digital pathology tools including machine learning algorithms, these digital platforms have the potential to enhance a pathologist's evaluation of histology in drug development and in the clinic. Utilization of these platforms allows pathologists to provide their results within hours of slide upload thereby shortening trial timelines, while allowing for accurate assessment consistent with the gold standard.

2.5.2 Analytical Validation Purpose and Methods

The Analytical Validation (AV) of AIM–NASH provides evidence of precision and accuracy in measuring each component (steatosis, lobular inflammation, hepatocellular ballooning) of the NAS grade and CRN fibrosis stage. AIM-NASH was assessed for non-inferior accuracy to individual manual readers (IMRs) scoring as compared to ground truth (GT) expert consensus. The ground truth was performed by two panels of 2 expert liver pathologists with a third pathologist serving as tiebreaker (Figure 1). In cases where the two primary readers disagreed with the score on any of the NASH components (steatosis, lobular inflammation, hepatocellular ballooning and/or fibrosis), the slide was sent out to the third tiebreaker pathologist. The tiebreaker pathologist was blinded to the scores of the two primary pathologists. If the tiebreaker pathologist agreed with one of the primary pathologists, this was then the final score for that NASH component. If the third pathologist disagreed with both primary pathologists, a joint panel call was held with the three pathologists to come to a consensus, with the tiebreaker providing the final score in the rare case that consensus was not reached. The tiebreaker pathologist was the same for both panels. Overall, 5 pathologists provided scores for ground truth. These 5 pathologists were unique and not used in any AI-assisted or for IMRs.

AIM-NASH was also evaluated for scanner repeatability and reproducibility performance in evaluating four histologic components - steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis. AV tests the performance of AIM-NASH itself, without the pathologist's final determination of scores.

Figure 1: Consensus Reads Establish Ground Truth



2.5.3 Data Sources:

AV of AIM-NASH was performed utilizing glass slides selected from 2 completed phase 2 (one non-cirrhotic and one cirrhotic) trials and one phase 3 NASH trial. One hundred and sixty-five (165) unique subjects were enrolled in this study from the non-cirrhotic phase 2 clinical trial dataset, 105 unique subjects were enrolled from the cirrhotic phase 2 clinical trial dataset, and 238 unique subjects were enrolled from the phase 3 dataset. For each of these trials, not all subjects had multiple time points (baseline and post-treatment) available for enrollment into the study. Overall, 322 samples were enrolled from baseline time points and 283 samples from post-treatment (both placebo and treatment arms) time points. Ultimately, 250 samples were utilized from a phase 3 trial, 217 samples from a non-cirrhotic phase 2 trial, and 139 samples from a cirrhotic phase 2 trial, with a total of 606 samples enrolled.

Table 8 provides the predefined target distribution of NASH scores enrolled into the AV study.

Table 8: Predefined Target Score Distribution for AV

Feature	Score	Percentage of Cases	
		Accuracy (n=600)	**Repeatability and Reproducibility** (n=150, subset of accuracy cases)
Steatosis	0	20%	20%
	1	20%	20%
	2	30%	30%
	3	30%	30%
Lobular inflammation	0	20%	20%
	1	20%	20%
	2	30%	30%
	3	30%	30%
Hepatocellular ballooning	0	20%	20%
	1	40%	40%
	2	40%	40%
Fibrosis	0	15%	15%
	1	15%	15%
	2	25%	25%
	3	25%	25%
	4	20%	20%

2.5.4 Major Findings and Conclusions

Non-inferior agreement (defined as greater than 0.1 less than the mean IMR concordance with GT) was achieved for each NASH histologic component and in certain subgroups relevant to clinical trial enrollment and efficacy endpoints (F4 vs other, F0+F1 vs other), and NAS $\geq$ 4 vs other). Furthermore, hepatocellular ballooning scores produced by AIM-NASH demonstrated significantly superior agreement to ground truth as compared to manual digital reads ($p<0.0001$).

100% repeatability was achieved when the algorithm was run on the same image multiple times at an external lab. When slides were scanned three times on the same scanner and by the same operator on different days, AIM-NASH met the target of >85% repeatability (agreement rates of 0.931, 0.963 0.958, 0.926, for steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis, respectively) and also displayed superiority over intra-reader percent agreement in the published literature 0.722, 0.553, 0.699, and 0.720 (Table 5) (14).

Reproducibility of algorithm performance on slides scanned at 3 different locations by 3 different operators narrowly did not meet the acceptance criteria (lower bound of the 95% CI is greater than 85%) for steatosis, lobular inflammation, and fibrosis (0.856, 0.847 and 0.868, respectively) but agreement was still greater than intra-reader agreement rates in published literature (0.633, 0.604, 0.509), respectively) (

Table 3) (14). AIM-NASH derived hepatocellular ballooning scores, however, did achieve statistical superiority over the performance goal of 85% (reported literature of 0.625) (14). The AIM-NASH data also showed superiority over the performance goal of 85% for clinical trial relevant histological categories of fibrosis low

(F0&F1), advanced fibrosis (F4), and NAS aggregate score ≥4. These data present AIM-NASH as an accurate and precise tool with strong potential for driving more rigorous and consistent clinical trial enrollment and efficacy analysis.

2.5.5 Overlay Validation Purpose and Methods

The AIM-NASH overlay validation is performed to validate the visual overlays that are displayed on the platform interface. Tissue detection models of AIM-NASH detect key histologic features (steatosis, lobular inflammation, and hepatocellular ballooning on H&E Whole Slide Images (WSIs) and fibrosis on trichrome WSIs) and are displayed as visual overlays in the viewing platform. These overlays are intended to facilitate the pathologist's review in the AIM-NASH scoring workflow. For this study, up to one hundred and sixty (160) 500 x 500-micron-sized frames were enrolled for each feature based on the enrolling expert NASH pathologist score, representing a wide range of disease activity. These frames were sampled from WSIs generated from NASH biopsy samples and the enrolled frames displayed the AIM-NASH model's predictions for steatosis, lobular inflammation, hepatocellular ballooning, and artifacts over the H&E-stained image, and fibrosis and artifacts over the trichrome-stained image in the form of a colored overlay. 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 under- or overestimating a given feature, defined as true positive (TP) and false positive (FP) success rates.

2.5.6 Data Sources:

The overlay validation slide set was sourced from the same population as AV and Clinical Validation (CV) studies and included 3 phase 2 studies (2 non-cirrhotic and 1 cirrhotic) and one phase 3 study (non-cirrhotic). Frames from 222 WSIs were enrolled. Overall, 312 unique H&E frames and 249 trichrome frames were enrolled.

2.5.7 Major Findings and Conclusions

The study assessed the accuracy of AIM-NASH overlays as a highlight tool for pathologists when reviewing the AIM-NASH generated scores. The results of this study show that the overlays are accurate when highlighting the features on NASH slides. Overlays for all features for TP and FP success rates met their acceptance criteria except for TP rate for ballooning where the lower bound of the 95% CI was slightly below the 0.85 acceptance criteria (0.833). Interestingly, pathologist A and B TP rates for hepatocellular ballooning were quite high, at 0.96 (95% CI, 0.911, 0.991) and 0.94 (95% CI, 0.89, 0988), respectively. However, pathologist C TP rate for hepatocellular ballooning overlay was only 0.72 (95% CI, 0.639, 0.805). This is not surprising as identification and quantification of hepatocellular ballooning cells has been shown to be difficult and inconsistent, even for experienced hepatopathologists. The significant variability in identification of frames and/or whole slides with ballooned cells highlight this variability, seen both in this study and amongst other expert CRN hepatopathologists in the literature (26). Overall, we conclude that the AIM-NASH overlay features are accurate in highlighting steatosis, lobular inflammation, hepatocellular ballooning, fibrosis, H&E artifact, and trichrome artifact as demonstrated by this overlay validation study.

2.5.8 Clinical Validation Purpose and Methods

The purpose of clinical validation (CV) is to measure the ability of AIM-NASH to assist pathologists in their assessment of NASH as the tool would be utilized in a therapeutic trial setting. The full AI-assisted workflow

includes pathologist review of the AIM-NASH scores for all 4 NASH components (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis), as well as review of sample, stain and scan adequacy. The study evaluated for non-inferior agreement of the AI-assisted scores with ground truth as compared to manual digital read agreement (from a minimum of three pathologists) with ground truth.

2.5.9 Data Sources

Two phase 2 studies (1 non-cirrhotic and 1 cirrhotic trial) and a subset of a phase 3 study (non-cirrhotic) were utilized for CV. From the phase 3 study, cases with two available time points (baseline and end-of-study) were randomly chosen based on AIM-NASH scores.

One hundred and fifty-four (154) unique subjects were enrolled in this study from the cirrhotic phase 2 clinical trial dataset and 470 unique subjects were enrolled from phase 3 clinical trial dataset. For each of these trials, not all subjects had multiple time points (baseline and post-treatment) available for enrollment into the study. Overall, 543 samples were enrolled from baseline time point and 435 samples from post-treatment time point. For the non-cirrhotic phase 2 trial, 523 unique samples were enrolled but the number of unique subjects is unknown as no time point information was available to the PathAI study team. Ultimately, 694 samples from phase 3 trial, 284 samples from the cirrhotic phase 2 study, and 523 samples from the non-cirrhotic phase 2 study were enrolled, with a total of 1501 samples.

2.5.10 Major Findings and Conclusions

The CV study demonstrates the accuracy of AI-assisted pathologist workflow in measuring each component of the NAS grade and CRN fibrosis stage in liver biopsies from patients screened and/or enrolled in a NASH clinical trial. CV was designed to test the performance of the pathologists aided by the AIM-NASH algorithm as it would be used in a NASH clinical trial setting, as the current standard of manual pathologist evaluation is characterized by significant inter- and intra-pathologist variability (6,8,14). Overall, non-inferior accuracy (defined as greater than 0.1 less than the mean IMR concordance with GT) was achieved for each NASH histologic component assessment (differences in WKs for steatosis 0.003, NI $p<0.0001$, lobular inflammation 0.123, NI $p<0.0001$, hepatocellular ballooning 0.15, NI $p<0.0001$ and fibrosis 0.008, NI $p<0.001$) with hepatocellular ballooning and lobular inflammation reaching superiority ($p<0001$). In histological categories relevant to clinical trial enrollment and endpoint assessments, AI-assisted pathologist scores also demonstrated improved accuracy as compared to manual digital scores, including significantly higher agreement to GT for assessment of samples as NAS$\geq$4 with a score of at least one for each component, as well as for NASH resolution (ballooning score of 0, lobular inflammation score of 0 or 1, and any value for steatosis), compared to unassisted reads. Additionally, given that the rejection rate for AI-assisted scores ranged from 0.37% to 1.83%, accuracy for AIM-NASH alone was also computed. Similar to AI-assisted results, AIM-NASH algorithm results alone demonstrate superior agreement with ground truth as compared to manual pathologist reads for lobular inflammation and hepatocellular ballooning ($p <0.0001$) and non-inferior agreement for steatosis and fibrosis (NI margin = 0.01, $p<0.0001$). Overall, the accuracy results from AIM-NASH alone and AI-assisted are very similar.

In conclusion, AIM-NASH assisted pathologists demonstrated non-inferior to superior accuracy to manual digital reads providing a solution for driving more standardized, rigorous, and consistent clinical trial enrollment and efficacy assessment.

2.6 Remaining gaps and a brief overview of how these will be addressed (if applicable)

If the context of use for the AIM-NASH algorithm has expanded beyond the use of only the Aperio AT2 scanner, a planned bridging study will be designed to address the gap that exists due to the AIM-NASH algorithm having been trained and validated on just one scanner at a single magnification. The objectives of this study are to assess AIM-NASH model performance and reproducibility between commonly used whole slide scanners. In this study, PathAI will assess the AIM-NASH scoring agreement between two magnifications (20x and 40x) at the case (one H&E and one trichrome slide) level by assessing the outputs (NAS score and CRN fibrosis stage, Table 1 and Table 2) of WSI generated at different magnifications. These same scoring agreements will be assessed between Leica's Aperio AT2 and Aperio GT450 scanners, Philips' Ultra Fast Scanner, and Ventana's DP200.

2.7 Conclusion

Pathologists aided by AIM-NASH are superior to unaided pathologists in accuracy in their assessment of hepatocellular ballooning, and lobular inflammation and non-inferior in their assessment of steatosis and fibrosis. These results are particularly impactful given the high degree of hepatopathologist disagreement in hepatocellular ballooning and lobular inflammation evaluation (26). Evaluation of repeatability and reproducibility of AIM-NASH outputs on WSIs scanned from Aperio AT2 scanners shows higher agreement than reported inter- and intra-reader agreement in literature (14). These data suggest that AIM-NASH reduces the impact of inter- and intra-reader variability in NASH clinical trial enrollment and endpoint measurement, enabling a more consistent and reliable assessment of therapeutics under development.

3 Need and Impact of Proposed Methodology and Methods of Measurement

3.1 Intended application of the novel methodologies and Context of use

Proposed Context of Use Statement

A monitoring biomarker as an adjunctive tool that aids the pathologist in assessing NASH disease activity (at baseline and subsequent time points) to produce the Nonalcoholic Fatty Liver disease activity score (NAS) components (i.e., steatosis, hepatocellular ballooning, lobular inflammation) and fibrosis stage in liver biopsies in pathologist-diagnosed NASH patients in NASH clinical trials.

Intended Patient Population

The biomarker is applicable to screening and at follow-up time points for phase 2 and phase 3 NASH trials. This includes patients with fibrosis stage ranging from 0-4 and NAS <4 and ≥4.

3.2 Integration of AIM-NASH into Drug Development

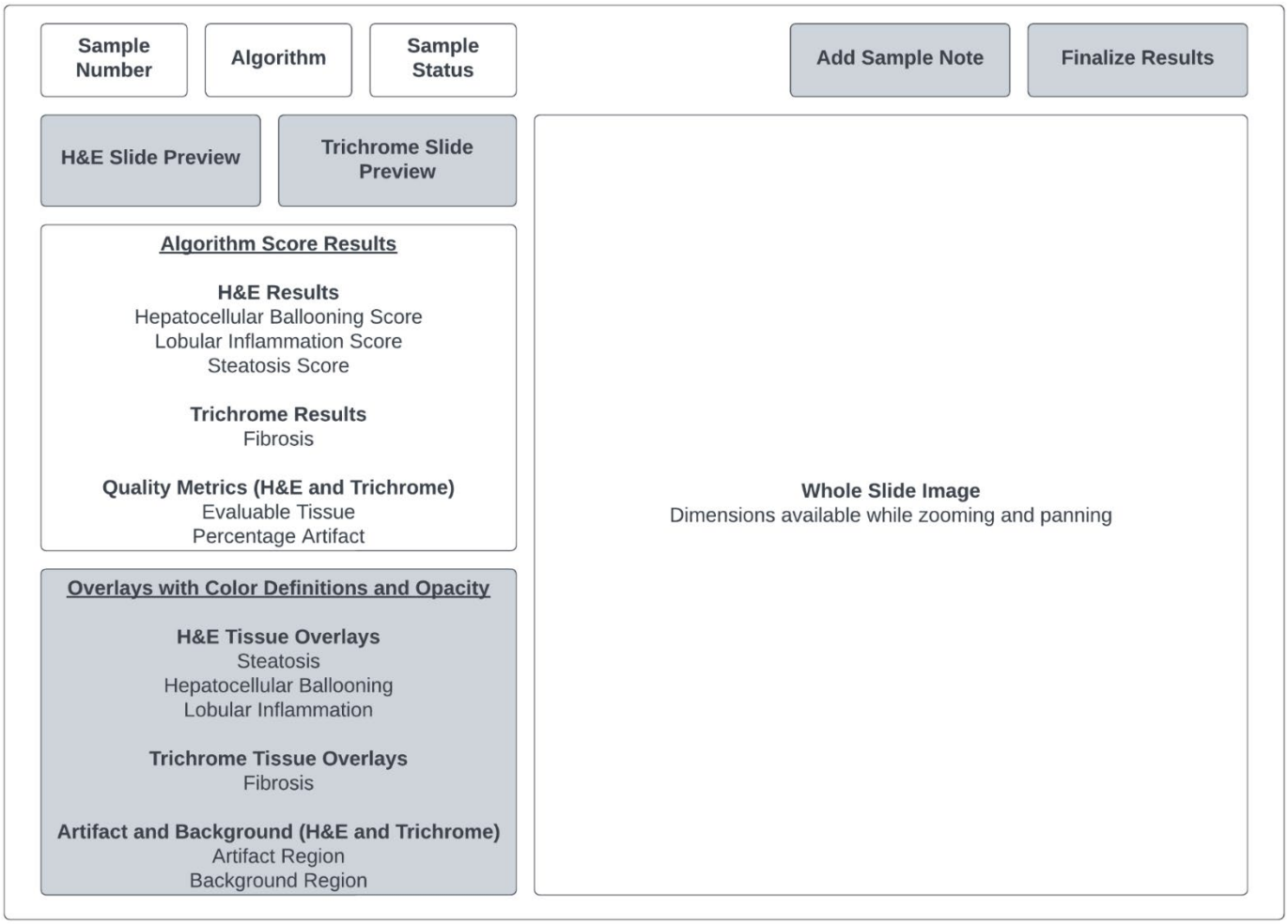
The AIM-NASH tool primary read-outs are the NASH CRN scores (NAS and CRN Fibrosis) which can be used for NASH clinical trial enrollment and assessment of histologic-based endpoints (steatosis,

ballooning, inflammation, and fibrosis stage). Although there is potentially inherent variability in the NAS-CRN score itself, it is the most commonly used NASH scoring system for clinical trials, and the AIM-NASH tool serves as a highly reproducible, consistent method to reduce said variability. Note that AIM-NASH will not make the determination on whether the study endpoint was met. Instead, endpoints will be assessed per study protocol by the sponsor, informed by AIM-NASH read-outs. See Figure 2 for a visual representation of the user interface a pathologist performs the AIM-NASH read-outs with.

Once the AIM-NASH algorithm is reviewed and qualified as a novel methodology, it will be deployed as a research tool to aid in drug development and in support of regulatory review. In our intended context of use, the patient care team remains in control of the patient diagnosis and care decisions, with the tool providing a suggested score to be reviewed and confirmed by the study pathologist.

Similar to our approach and understanding from the US FDA, once a determination has been made that AIM-NASH and the CTS platform can be relied upon for drug development applications and the tool is qualified as a novel methodology, no additional conformity assessment is planned.

Figure 2: Pathologist User Interface for Read-outs



To determine the impact of using AIM-NASH in a NASH clinical trial, we interviewed subject matter experts, reviewed trial protocols, and conferred with NASH trial sponsors. Our research determined that the integration of AIM-NASH in a trial would not impact the overall trial workflow significantly (Figure 3 and Figure 4), nor introduce any new risk which could affect/change current patient standard-of-care (SOC).

Figure 3: Current NASH Clinical Trial Workflow with Manual Pathology Review

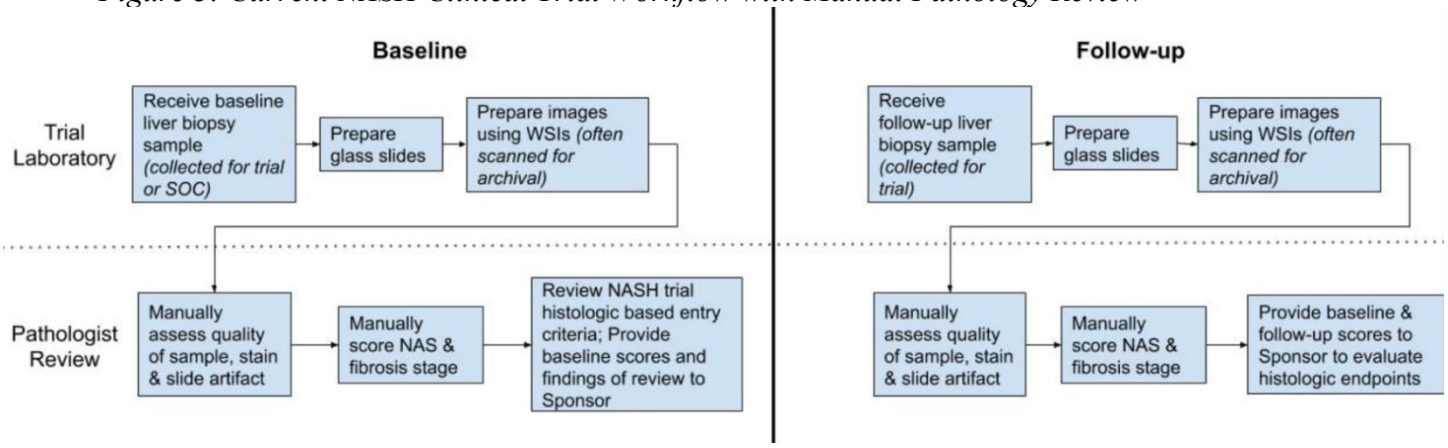
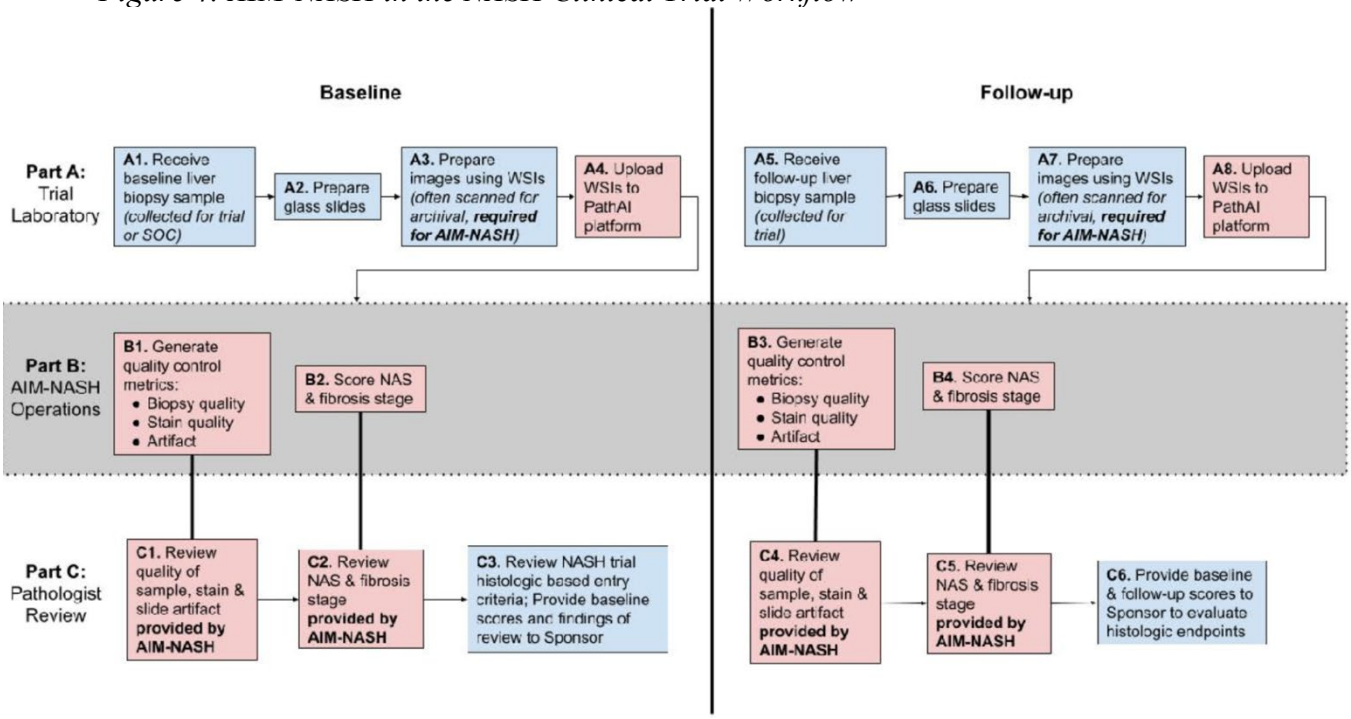


Figure 4 indicates, in red, where new steps are added to the clinical trial workflow where AIM-NASH is implemented, and the below text refers to this workflow.

Figure 4: AIM-NASH in the NASH Clinical Trial Workflow



Staff and Setting

Trial Laboratory: The trial workflow applies to local or central laboratories. AIM-NASH design considered the need to develop a tool that could perform well with staining variability within and between labs, with assurance of quality practices at the site level. Samples must be prepared, stained and scanned in a College of American Pathologists/ Clinical Laboratory Improvement Amendments (CAP/CLIA)

certified (or EU equivalent) laboratory that has calibrated equipment and a quality system in place. Given that the biomarker uses a cloud-based platform, this may be used by a lab in any location with connectivity. CAP/CLIA certification includes qualification, validation, and regular proficiency testing of lab equipment and laboratory staff. Additionally, PathAI development is performed in compliance with a Quality Manual (that requires a regular internal audits and management review) and the oversight of a qualified individual responsible for the implementation and maintenance of quality activities related to laboratory testing. These controls are equivalent to practices in the EU and pathology labs located in the US or EU can be included in the Qualification exercise.

Trial Laboratory Staff: Trial Laboratory staff will adhere to the Instructions for Use (IFU) for the scanner being used and the AIM-NASH tool.

Pathologist: The pathologist may be a local pathologist or central reader. The pathologist will be trained to evaluate the AIM-NASH results via a PathAI developed webinar. Pathologists should gain comfort with the system on example cases prior to performing analysis of clinical trial samples. Records of live session or webinar training may be trial-specific and will be maintained in PathAI QMS. Study specific proficiency testing should be maintained per sponsor requirements within the trial master file.

Figure 3 shows high-level current, traditional workflow with manual pathology review (which may be performed using glass slides or digitized whole slide images). Pathologists are blinded to time point and screen failures can be spiked in regular intervals to prevent temporal bias; the full workflow is agreed upon with the Sponsor and documented in the central histology manual.

Trial Workflow (Figure 4)

Baseline Timepoint

• Part A: Trial Laboratory (applicable whether a local site laboratory or central laboratory is utilized)

- A1: As is specified in the trial protocol, biopsy is either obtained from standard of care procedure (from patient care team, max age of biopsy must meet trial specified protocol criteria) with suspected NASH, or biopsy is performed for the trial. Note: If a historical biopsy is used, either the block or unstained slides may be sent to the trial laboratory depending on quantity of tissue available, or WSI images from local laboratory can be sent directly to trial pathologist, which may be preferable to conserve tissue (if local laboratory meets the requirements for and has qualified scanner in place).
- A2: The trial laboratory prepares and stains the patients' screening biopsy sections using the appropriate laboratory procedures for H&E and Masson trichrome staining.
- A3: The laboratory scans each glass slide at 40x using a calibrated, validated WSI scanner in accordance with the lab's quality procedures, the scanner IFU, and any other requirements outlined in the AIM-NASH qualification process. The WSI is checked for quality by the scanner technician and rescanned if needed until acceptable quality is achieved.
- A4: The histotechnologist or appropriate laboratory personnel then uploads the slides using a secure login to the PathAI platform and the trial pathologist is notified when cases are ready for review.

• Part B: AIM-NASH Operations (via a secure, validated PathAI cloud-based system)

- B1: Generate NAS grades and CRN fibrosis stage.

• Part C: Pathologist Review (high level workflow; role of the pathologist is described below in depth in Figure 4)

- C1: Sample, stain, and slide quality assessment is performed to ensure the tissue is indeed liver, assess biopsy adequacy and confirm stain quality and artifact level.
- C2: Review NAS component and CRN fibrosis stage provided by AIM-NASH algorithm. Either accept score or reject by 2 or more scores for any component and send for consensus review. See Figure 5 and Figure 6 for specific criteria and process. Determine final score.
- C3: Review for any other relevant histologic findings (per trial inclusion/exclusion criteria) and integrate with score assessment to determine whether a patient meets histologic-based inclusion criteria for the trial (send findings to sponsor data management).

• Notes

- We recommend for best practice, and as is outlined in the trial histology manual, trial pathologists are blinded to patient data, except for accession number, and do not make diagnostic claims, but only evaluate for histologic-based inclusion criteria.
- Incidental findings during review of the biopsy will be handled according to trial protocol. This may include communication (and shipment of the slides) to the referring physician. Use of AIM-NASH will not interfere with the handling of incidental findings. Use of AIM-NASH in the trial will not affect this communication either way.

Follow-up Timepoints

● Part A: Trial Laboratory
- Workflow mirrors enrollment sample processing, but this sample is collected for trial purposes only (A5)
- Workflow mirrors enrollment sample processing

● Part B: AIM-NASH Operations
- Workflow mirrors enrollment sample processing

● Part C: Pathologist Review
- Workflow mirrors enrollment sample processing
- The sponsor will determine if the histologic change meets the endpoint criteria per trial protocol

Note on baseline vs. follow-up biopsy quality: During research into NASH trial protocols, and in discussion with labs, blocks or slides are sent for sectioning or staining at a central lab for the majority of trials with histologic-based enrollment/endpoints (focusing on phase 2 and 3 trials). However, this biomarker was developed to accurately score in the event of staining variability. Pathology labs communicated that sectioning new slides from historical patient biopsy blocks can be less than ideal (often there's not a lot of tissue left). If a scanner is validated under the biomarker COU, we propose that the original slides are QC'd for biopsy and staining quality and then scanned at the local site if feasible, to mitigate the risks that already occur with inadequate tissue being left in the block that is sent to the central lab for the trial. Regardless, we will stratify results, where information is available, from Analytical Validation (AV) and Clinical Validation (CV) to determine and ensure accuracy and reproducibility of baseline screening biopsies, separately from follow-up biopsies.

We are confident that the use of AIM-NASH would not introduce additional bias in the trial and instead that it would mitigate and lessen some of the noise that currently impedes accurate histologic measurement.

Role of the Pathologist

As we have described in the above clinical workflow, the pathologist first reviews sample adequacy based on trial protocol and if needed, can request a re-stain or a rescan of the slide. If the sample is deemed adequate, the pathologist then reviews the output of AIM-NASH and may also make note of any unexpected findings.

The decision tree in Figure 4 shows the discrete steps taken by the pathologist during this interaction with the biomarker. As shown in Figure 4, the pathologist will assess sample quality according to the clinical trial protocol. If the sample is acceptable, the pathologist will review the H&E based NAS component scores generated by AIM-NASH. If the pathologist accepts these scores (within +/- 1 point per individual feature), they will then review the AIM-NASH CRN fibrosis stage. If the pathologist finds this score acceptable (also within +/- 1 point), they will record their agreement and sign-out the case.

We and others have provided extensive evidence that NASH scoring using the CRN system is prone to wide ranges of variability both intra-and inter-rater (14,22), even amongst the CRN pathologists. Additionally, there are varying interpretations in identifying the relevant histologic features themselves, even before any quantification for scoring, especially for hepatocellular ballooning (26). The 2-point rejection workflow is intended to reduce individual bias and inconsistency, a widely understood and documented challenge associated with NASH scoring, while still allowing the pathologist to reject if sample quality or evaluability is not acceptable (either on the biopsy, staining, or scanning level), or if there is additional pathology present. Additionally, algorithm overlays can be toggled off and on to facilitate review based on pathologist preference.

The pathologist workflow for reviewing AIM-NASH results involves sample quality assessment, review of NAS AIM-NASH grades and review of trichrome fibrosis AIM-NASH stage. The pathologist may choose to agree with all results and release the score or take actions in accordance with sample quality failure or follow the algorithm result rejection workflow (Figure 6). At each of these steps, the pathologist makes decisions in accordance with the AIM-NASH guidelines and clinical study protocol. Based on their review of the WSIs, the pathologist has the flexibility to reject a sample based on quality or disagree with the AIM-NASH scoring of the H&E or trichrome images if the feature score is 2 or more points off their assessment. This workflow for the pathologist is consistent with that in current NASH clinical trials.

As Brunt points out in her 2020 editorial, pathologists' interpretation of these slides goes beyond providing a score, and the pathologists will still be performing this assessment manually, using the PathAI validated viewing platform (27). The decision tree in Figure 5 shows the discrete steps taken by the pathologist during this interaction with the biomarker. This proposed workflow was developed with subject matter experts and based on research into existing practices. The key takeaway is that the pathologist will play an active role in the interpretation of AIM-NASH results and will accept or reject the scores (Figure 5 and Figure 6).

Figure 5: Pathologist Review Workflow and Inputs into Decision

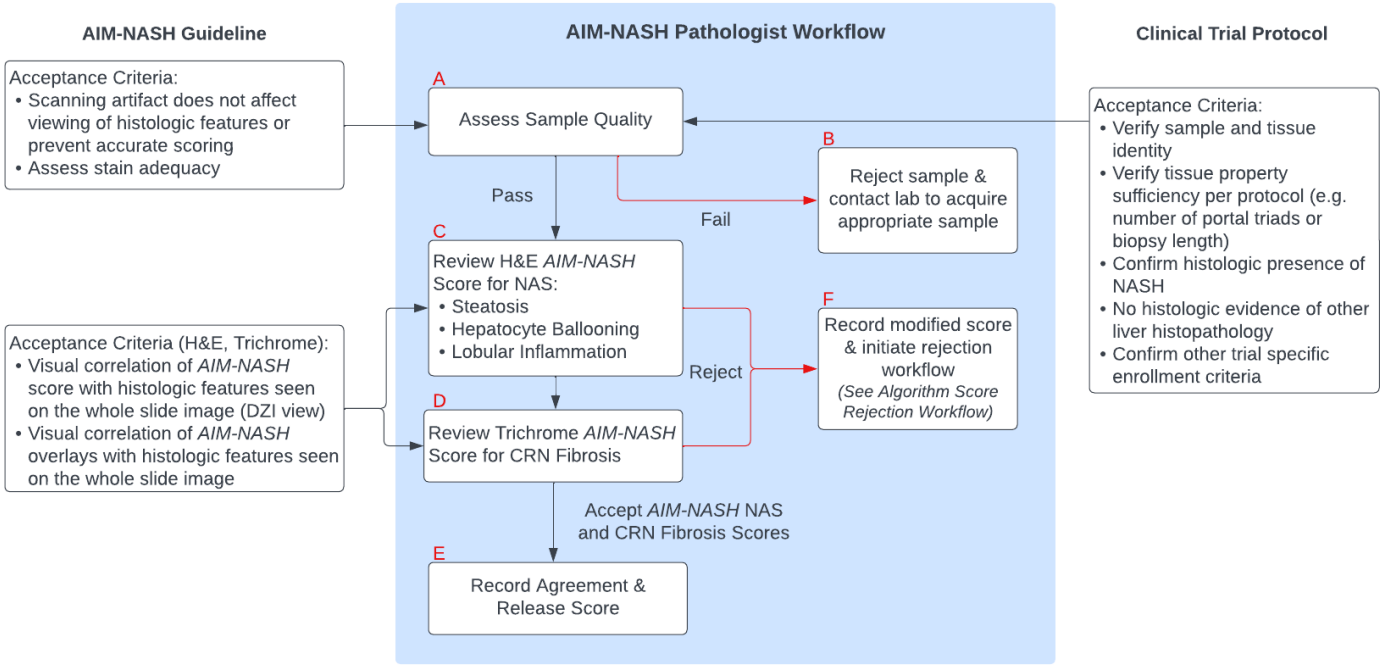
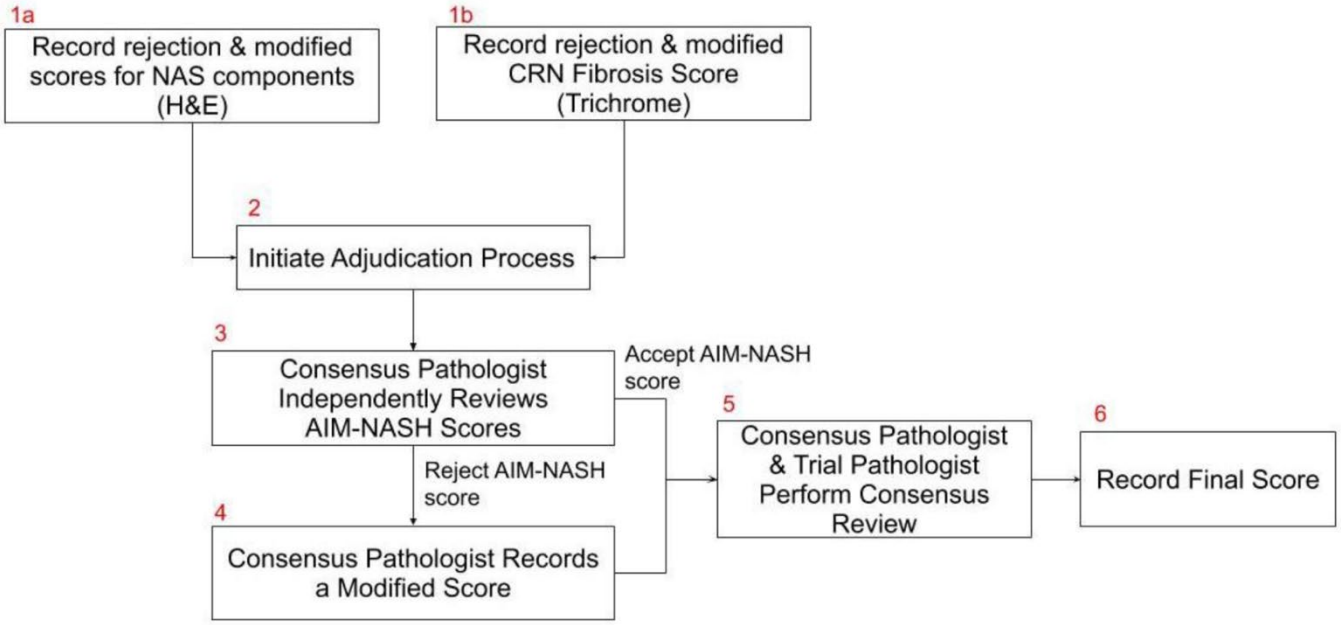


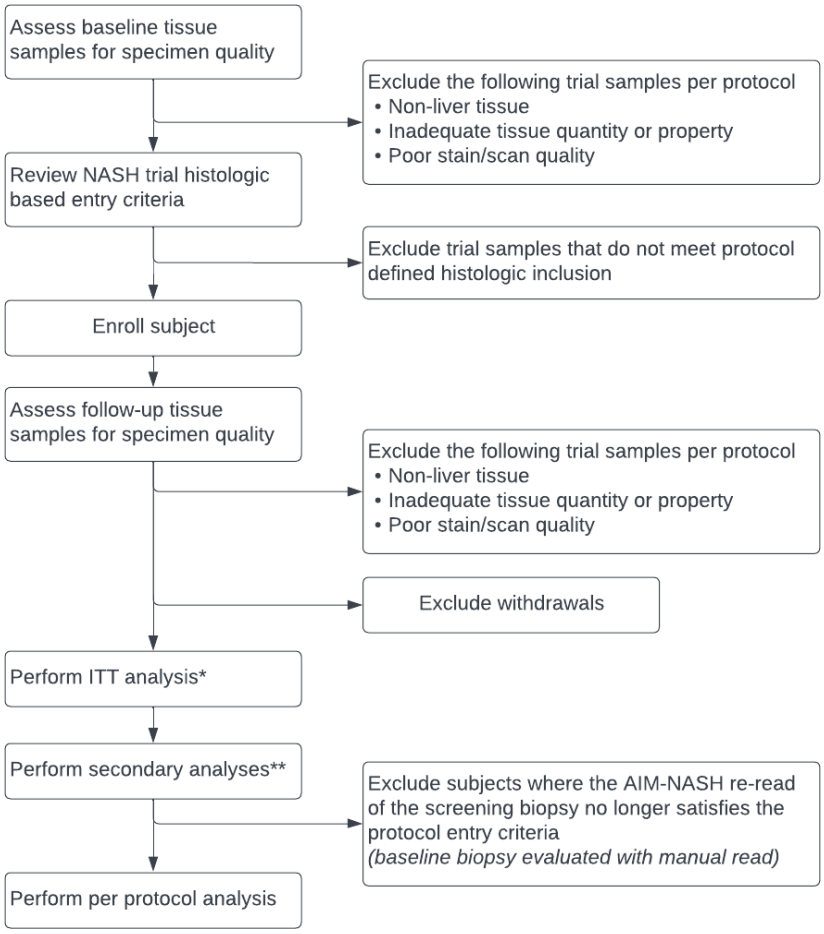
Figure 6: Algorithm Score Rejection Workflow



Post-enrollment Workflow

AIM-NASH assisted reads may also be utilized post enrollment with the proposed workflow detailed in Figure 7. For an Intent-to Treat analysis, in trials where AIM-NASH is implemented after the first subject is enrolled and baseline reads where performed manually, the pathologist may utilize the AIM-NASH to re-analyze those samples. AIM-NASH may also be utilized for secondary analyses where baseline reads were done manually. The pathologist may re-analyze both baseline and follow-up samples utilizing the AIM-NASH tool. The secondary analysis will use the reconciled results and exclude patients who no longer meet the score-based entry criteria.

Figure 7: Incorporation of AIM-NASH Post Enrollment



Risks of AIM-NASH are associated with failure of the tool to perform as expected, leading to incorrect test results. PathAI will minimize the risk from incorrect results by performing tests, to optimize and validate the device (see Analytical and Clinical Validation Plan in section 4).

3.3 The limitations to the qualification sought

AIM-NASH is intended as a supplement to pathologist review and is not a substitute. The tool should only be used in tandem with the assessment of a qualified liver pathologist.

3.4 Potential impact of AIM-NASH on current regulatory guidelines

The AIM-NASH methodology does not change the current recommended trial inclusion criteria or endpoints but aims to provide a more standardized and precise method of measuring them. The EMA reflection paper also states that "Liver biopsy and histology have been widely criticized for sampling error and intra- and inter-observer variability. This proposed novel methodology is intended to address the latter part of this unmet need.

3.5 Setting in which the AIM-NASH tool will be applied

AIM-NASH is intended to be utilized in NASH clinical trials at the time of enrollment and follow-up. Typically, in NASH clinical trials, a follow-up biopsy is collected, often at 52 weeks, and the above NAS and fibrosis scoring systems are again assessed with the objective of capturing change in disease state to assess whether endpoints have been met. For NASH trials, EMA has suggested to include patients with definite NASH, compensated NASH or decompensated NASH based on NAS and fibrosis staging (13).

While the progression of NASH to various stages of fibrosis is not fully understood, patients with stage 2-3 fibrosis are at higher risk for progression to cirrhosis within 10 years, as well as having an increased mortality risk (28). The features contributing to the NAS and fibrosis scores will be evaluated specifically, to determine whether any of the following intermediate endpoints from the EMA guidance have been met:

Stage 2 & 3 fibrosis *(demonstrated in co-primary fashion)*

1. The resolution of NASH – with the presence of any grade of steatosis, no ballooning, and only minimal (grade 1) lobular inflammation and – at the same time – no worsening of the stage of fibrosis.
2. The improvement of fibrosis by at least 1 stage without any worsening of NASH (no worsening of ballooning and lobular inflammation, a 1 grade change in steatosis may be acceptable).

Stage 4 fibrosis

Improvement of liver cirrhosis to non-cirrhotic liver disease (1 or more-point improvement in fibrosis stage)

Population for use:

- Adults, 18 years and older.
- Patients enrolled in a NASH clinical trial or screened for trial entry, with a NASH confirmed biopsy or biochemical criteria and/or imaging evidence of steatosis/steatohepatitis/fibrosis in addition to known risk factors for NASH.

Whole Slide Image Considerations for Clinical Trial Use:

- Formalin-fixed, paraffin-embedded (FFPE) liver biopsy tissue should be stained with H&E and trichrome according to the package insert.

- H&E and trichrome-stained slides should be scanned by CRO on validated scanner(s) and the WSIs should be quality checked according to their instructions for use.

3.6 Currently available tools in patient care and drug development:

Accurate scoring of NASH during a clinical trial is required to measure histologic change and evaluate trial endpoints. Given the unmet medical need, it is important to identify and gather evidence for reliable endpoint measurements that will help to accelerate drug development for both non- cirrhotic and cirrhotic NASH patients. Accurate, precise, reproducible, and easy to implement histologic-based serial measurements used during trials will help to evaluate whether patients are responding to and will likely benefit from a therapy when clinical outcome is determined at later time points.

NASH disease activity is assessed histologically in the clinical trial setting by the non-alcoholic fatty liver disease (NAFLD) Activity Score (NAS) and the presence of steatohepatitis (8–10,17,28). The semi-quantitative and subjective nature of these manual scoring methods may be part of the root cause of discordance. In addition to the risk of inter- and intra-reader variability, NASH trials are subject to systematic bias in the form of temporal bias. Methods used to guard against systematic bias include randomization of slide assignment (mixing baseline and follow-up samples) and assignment of multiple central pathologists. However, the possibility of systematic bias is more difficult to control when grading and staging are required for trial entry.

Reference Standard: Given the issues with inter- and intra-rater reliability described earlier in this submission, we will not rely on the score of one pathologist to provide the reference standard for model evaluation and will instead use a consensus from multiple pathologists.

3.7 Technical Considerations and Characteristics of the Biomarker

Biomarker Name and Type

- Biomarker Name: AI-based Histologic measurement of NASH (AIM-NASH)
- Type of Biomarker: Histologic based, imaging modality, measurement based on machine learning
- BEST Classification: Monitoring biomarker

3.7.1 Technical Aspects Summary of the Biomarker:

Biomarker Source

The source of the biomarker is FFPE liver biopsy tissue. The tissue is collected at time of screening (or within an approved window, per trial protocol) and during follow-up timepoints for enrolled patients. Masson trichrome stained slides are required for fibrosis staging and H&E staining is used for NAS components (Table 1 and Table 2). Glass slides are scanned at a resolution of 40X and the histotechnologist scanning the slides performs quality control on the scanner according to machine specific instructions.

The biomarker results are composite measures. NAS consists of the independent lobular inflammation, steatosis, and hepatocellular ballooning scores per the NASH CRN histologic scoring system. Fibrosis stage consists of multiple biologic entities measured separately, such as trichrome staining of large septae, pathological periportal fibrosis, and fibrotic septa, but with a single output score (fibrosis stage).

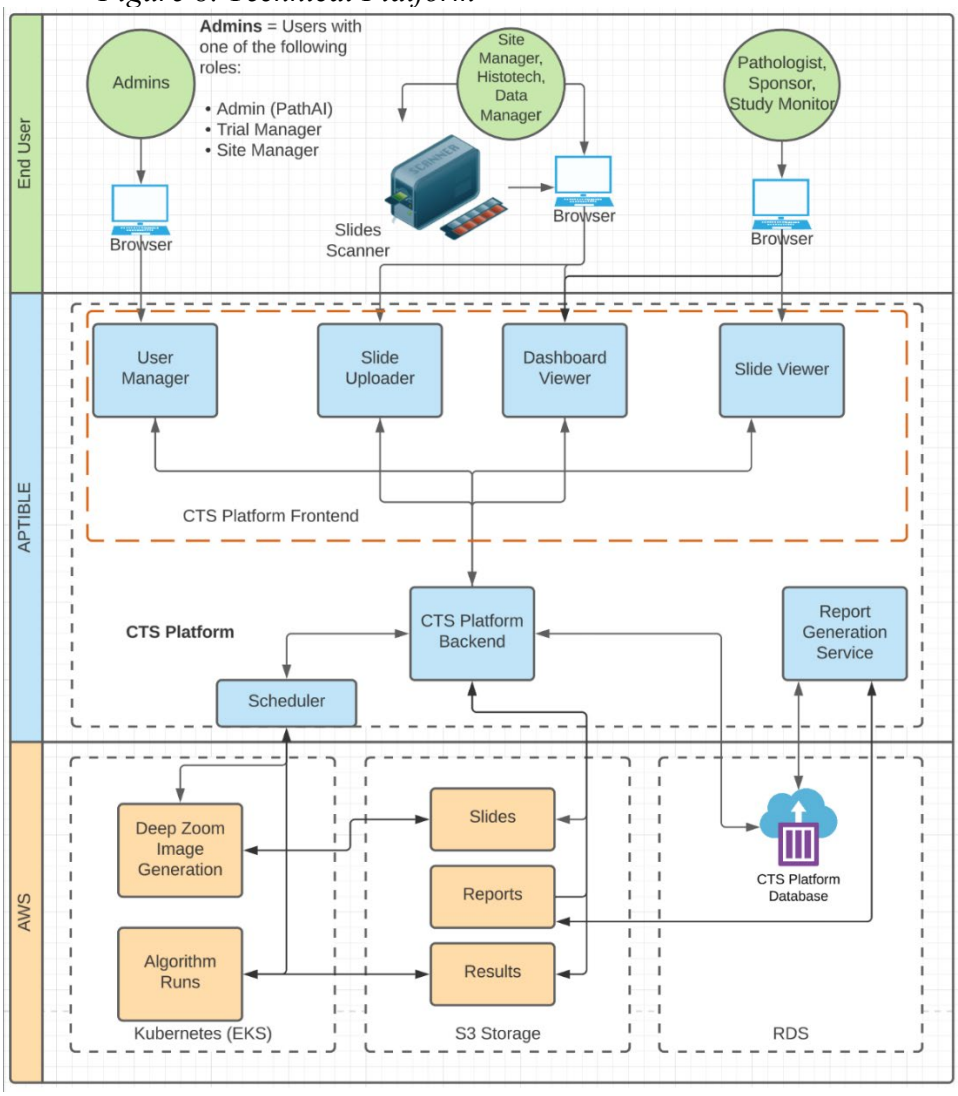
Technical Platform

The technical platform used to measure the biomarker consists of the following elements (Figure 8):

1. Qualified WSI Scanner at Trial Site: Slides must be scanned at a CAP/CLIA (or European equivalent, ISO 15189) compliant laboratory with the validated Aperio AT2 scanner at 40X magnification. The histotechnologist scanning the slides performs quality control on the scanner according to machine specific instructions.

2. AISight Clinical Trials Platform: The AIM-NASH algorithm is hosted on the AISight Clinical Trials Platform developed by PathAI. The Platform serves as an interface for viewing whole slide images and algorithm outputs. The platform is version-controlled, and release notes will be maintained any time there are patches or fixes to the platform. PathAI provides an Instructions for Use (IFU) that details use of the tool.

Figure 8: Technical Platform



Workflow: A histotechnician or a lab assistant at a clinical trial site or central laboratory uploads slides into the AISight Clinical Trials Platform. The data uploaded will include an anonymized Subject level identifier, Accession level identifier (capturing the time point at which patient's slides were prepared) H&E slide, and trichrome Slide. The upload process could happen in one of two ways:

- The user may use the direct web uploader in the Platform user interface which allows them to manually select and upload WSI files and manually enter subject metadata.
- The user may use our bulk ingestion mechanism which allows them to drop the whole slide image files with an associated manifest (in csv format) containing metadata in a pre-designated Simple Storage Service (S3) bucket on Amazon Web Services (AWS).

Upon completion of upload, the H&E and trichrome images for each subject are ingested by the platform and the AIM-NASH algorithm run is automatically initiated. When the algorithm run has been completed, the AISight Clinical Trials Platform displays the algorithm results which include:

- Hepatocellular ballooning Grade (H&E Slide)
- Steatosis Grade (H&E Slide)
- Lobular Inflammation Grade (H&E Slide)
- Fibrosis Stage (trichrome Slide)

In cases where an input slide is insufficient due to trial specific criteria or AIM-NASH biomarker guidance, users have the ability to replace one or both input slides and re-send for algorithm analysis.

In cases where extraneous tissue is present on the slide, the pathologist will have the ability to exclude this from scoring using an exclusion annotation tool and have the algorithm analysis be updated to exclude this region from analysis.

- The Platform allows the pathologist to record their review and modify each score produced by the AIM-NASH algorithm if they disagree by 2 or more for each component of the NAS and CRN fibrosis scores.
- The Platform will automatically determine whether a consensus review between the primary reviewing pathologist and the secondary pathologist is required. The platform enables both pathologists to review the case and record results from the consensus review.
- Upon release of case scores, the Platform generates a report for each accession.

AWS Infrastructure

- Versioned images of the Platform's front end user interface (UI) and backend services will be deployed onto AWS infrastructure using docker application containers managed by PathAI.
- S3 will be utilized for bulk image ingestion, and by the Platform to store image data [WSI and deep zoom images (DZI)] and algorithm results. The bulk image ingestion/slide transfer process has been validated for use with the PathAI tool.
- AWS Identity and Access Management (IAM) is user and permission management for AWS resources. Cognito is also used if SSO authentication is used.

- Additional containers in Elastic Kubernetes Service (EKS) will be used for logging and monitoring services.
- PostgreSQL using RDS and SQS containers will be provisioned and used as data stores for the Platform.
- The Scheduler within the Platform will utilize EKS to provision (EC2) instances to be used by the algorithm to run based on specifications provided by the algorithm.

Machine Learning Pipelines

Image data (WSI) is fed into a sequence of appropriate deep learning models and image processing tasks to generate segmentation overlays, NAS component scores, and fibrosis scores.

Display: Labs should follow their own quality management system and it is recommended that before trial start, that the pathologist validates or gains comfort by comparing review of glass slides with AIM-NASH outputs for a small set of glass slides (29). Of note, in this qualification package, evidence of the validation of the AISight Clinical Trials platform and AISight Translational platform are provided.

Information Security Management System (ISMS)

The technical platform development and deployment falls under PathAI's ISMS that is intended to address and comply with (EU) 2016/679 (GDPR) that requires data to be processed according to seven protection and accountability principles as outlined in Article 5.1-2 (Lawfulness, fairness, and transparency; Purpose limitation; Data minimization' Accuracy; Storage limitation; Integrity and confidentiality; and Accountability). The ISMS serves as PathAI's GDPR policies.

PathAI's GDPR policies apply to all of PathAI's processing of personal data, and covers all of Path AI's information systems, workforce members, and contractors, For the purposes of the ISMS, "'personal data" means any information relating to an identified or identifiable natural person in the European Union (EU), regardless of where the collection occurs. Measures implemented to enable these principles are documented and traced.

3.7.2 Biomarker Measurement Process

The biomarkers are measured using ML models that detect and score NASH histologic features. The primary read-outs of the AIM-NASH ML model pipeline for each case (single biopsy with H&E and trichrome WSIs) are the NASH CRN ordinal scores for steatosis, hepatocellular ballooning, lobular inflammation grade and fibrosis stage, following the NASH CRN histologic scoring system (Table 1 and Table 2). The histologic-based efficacy endpoints recommended by EMA can be computed directly using the AIM-NASH primary read-outs on baseline and follow-up biopsy pairs.

Table 9 summarizes the purpose, model input, and model output of AIM-NASH ML models, Models 1-5, during inference. Outputs from Models 3a and 3b are concatenated to create a single output from Model 3. Similarly, outputs from Models 4a, 4b, and 4c are concatenated to create a single output from Model 4. The pipeline of AIM-NASH ML models is illustrated in Figure 9 and Figure 10, demonstrating the models that are run in sequence to generate the NAS and fibrosis scores, respectively, during inference. To summarize, if a region is classified as artifact or slide background by Model 1 and as a feature by Models 2 or 3, the final assignment will always be artifact or slide background. Specifically, on a new H&E image the Artifact model (Model 1) and H&E tissue model (Model 2) are run in parallel on the image. The artifact model produces predictions for classifying pixels

as Usable Tissue, Artifact or Slide Background. The artifact model predictions are then used to mask Artifact and Background pixels from the outputs of Model 2. The masked output of Model 2 is used to create the input to Model 4. For trichrome, a similar process is followed where Models 1, 3a, and 3b are run in parallel, the Artifact model is used to mask the outputs of Models 3a and 3b, and the masked output of Model 3 is used as input to Model 5.

The flow of information in the pipeline of ML models is as follows:

Step 1: Collect exclusion annotations, WSIs, and employ convolutional neural networks (CNNs) to generate overlays of relevant features.

(A) H&E and trichrome WSIs are submitted for scoring.
(B) An artifact detection model (Model 1) detects and quantifies artifacts on H&E and trichrome the WSIs; regions of artifact are excluded from subsequent evaluation and scoring.
(C) Tissue detection models detect the relevant histologic features on H&E WSIs (Model 2) and trichrome
(D) WSIs (Model 3); the detected features are visualized as WSI overlays (described in detail below).

Step 2: Feature overlays, plus expert liver pathologist slide level scores serve as input to Graph Neural Networks (GNNs) to generate slide level scores based on and validated against NASH CRN scoring system.

NASH CRN scoring models use the detected histologic features as input (in addition to whole slide image scores supplied from multiple pathologists) to score steatosis, ballooning, and lobular inflammation on H&E WSIs (Model 4) and fibrosis on trichrome WSIs (Model 5).

The final read-outs of the ML model pipeline are the NASH CRN scores.

Table 9: Description of AIM-NASH ML Models

Model Name	**Purpose**	**Model Input (during inference)**	**Model Output**
Segmentation Models			
Model 1: Artifact Model	To remove unwanted regions which should not even be considered as input for other models	H&E Whole Slide Image	Output is a 3D matrix of class probabilities for each pixel in the WSI. Artifact segmentation classes: Usable tissue, Artifact area, and Background area.
Model 2: H&E Tissue Model	Detect steatosis, hepatocellular ballooning, lobular inflammation regions and other to compute features of interest	H&E Whole Slide Image	Output is a 3D matrix of class probabilities for each pixel in the WSI. H&E segmentation classes: Lobular Inflammation, Portal Inflammation, Interface Hepatitis, Bile Duct, Blood Vessel, Normal Hepatocytes, Hepatocellular Swelling, Hepatocellular Ballooning, Steatosis, Microvesicular Steatosis, Normal Interface, and Other/remaining Tissue.

Model Name	Purpose	Model Input (during inference)	Model Output
Model 3a: trichrome Tissue Model	Detect fibrosis region to compute features of interest	trichrome Whole Slide Image	Output is a 3D matrix of class probabilities for each pixel in the WSI. trichrome segmentation classes: Collagen/Fibrosis, Bile Duct, Lumen, Blood Vessel, and Other/remaining Tissue
Model 3b: trichrome Pathological Fibrosis Model	Detect pathological fibrosis region in all fibrosis area	trichrome Whole Slide Image	Output is a 3D matrix of class probabilities for each pixel in the WSI. trichrome Pathological Fibrosis segmentation classes: Pathological Fibrosis, Normal Collagen and Other/remaining Tissue
Graph Neural Networks			
Model 4a: H&E GNN Model - Steatosis	Compute slide-level NAS ordinal scores for Steatosis	Raw model output from Model 2	Output is a single integer score for Steatosis (values in the range of 0-3)
Model 4b: H&E GNN Model - Ballooning	Compute slide-level NAS ordinal scores for Hepatocellular Ballooning	Raw model output from Model 2	Output is a single integer score for Hepatocellular Ballooning (values in the range of 0-2)
Model 4c: H&E GNN Model - Lobular Inflammation	Compute slide-level NAS ordinal scores for Lobular Inflammation	Raw model output from Model 2	Output is a single integer score for Lobular Inflammation (values in the range of 0-3)
Model 5: trichrome GNN Model	Compute slide-level CRN ordinal score	Raw model outputs from Model 3a and Model 3b	Output is a single integer score for CRN score (values in the range of 0-4)

Figure 9:H&E Inference Pipeline

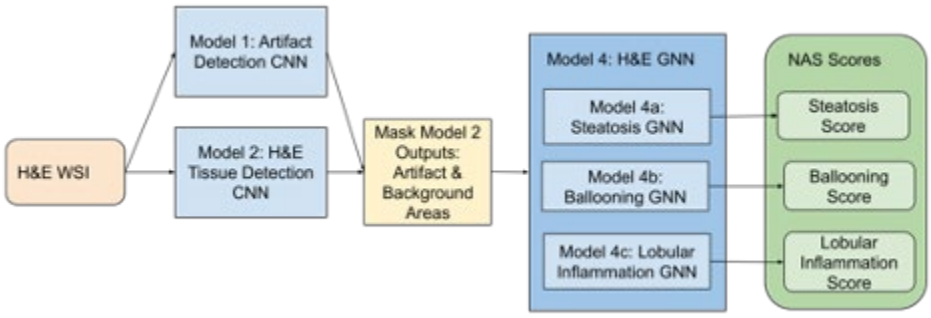
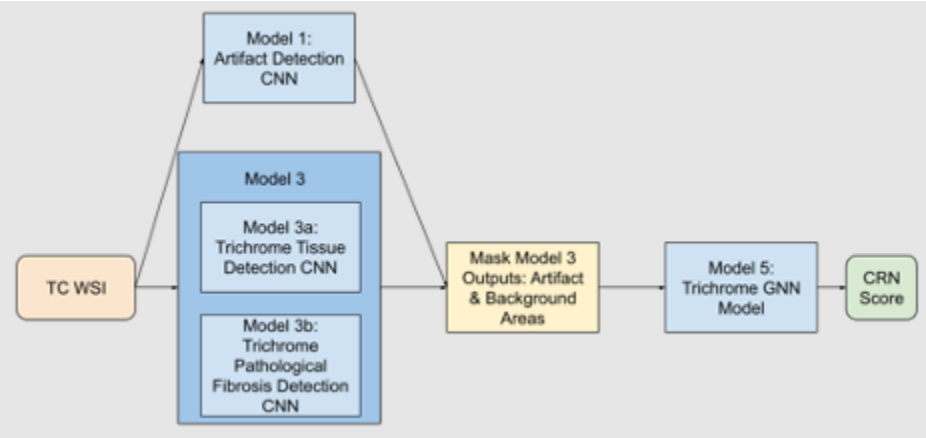


Figure 10: Trichrome Inference Pipeline



If the whole slide or a significant portion is classified as artifact such that each cluster used for graph construction contains less than 10 tissue pixels, no graph is constructed and the GNN models return NA as a score for the NAS components (if the slide is stained with H&E) or the CRN score (if the slide is stained with trichrome). This means the model is unable to compute the score for that slide. The algorithm does not automatically reject slides based on the amount of usable tissue or artifact except the above explanation around GNN. Pathologists determine the quality of composition of NASH tissue.

General Characteristics of WSI Dataset for AIM-NASH ML Model Development

The pipeline of ML models in AIM-NASH was developed using WSIs from multiple liver clinical trials, comprising a wide range of drug classes and disease severity as shown in Table 10. Specifically, the datasets used for model training included H&E and trichrome WSIs from multiple NASH clinical trials, as well as from liver biopsies of primary sclerosing cholangitis (PSC) and chronic hepatitis B infection (HBV) clinical trials. These data allowed the models to learn histologic features that are not characteristic of NASH but could present on liver biopsies (e.g., interface hepatitis) (17) enabling more precise identification of NASH-specific histology.

Table 10: Overview of Available Datasets for Developing ML-Based Image Segmentation and CRN Scoring Models

Clinical Trial	Phase	Total Available Sample Size	Drug Class	Enrollment Criteria
Training Datasets				
NASH Training Datasets				
1	3	H & E: 2188, trichrome: 2188	ASK1 inhibitor	NASH diagnosis; Fibrosis F3
2	3	H & E: 2488, trichrome: 2478	ASK1 inhibitor	NASH diagnosis; Fibrosis F4
3	2B	H & E: 528, trichrome: 528	Monoclonal antibody directed against LOXL2	NASH defined as steatosis > 5% w/ associated lob inflammation: Ishak stage 3,4
4	2B	H & E: 561, trichrome: 554	Monoclonal antibody directed against LOXL2	NASH diagnosis; Ishak stage 5,6
5	2	H & E: 158, trichrome: 163	ASK1 Inhibitor, monoclonal antibody directed against LOXL2	Evidence of NASH w/ fibrosis on biopsy
6	2	H & E: 304, trichrome:304	PPARδ agonist	Definite NASH; NAS≥4 w/ 1 per component; Fibrosis F1, F2, F3
Non-NASH Training Datasets				
7 & 8	3	H & E: 2181, trichrome: 1104	Nucleotide analogue (antiviral)	HBV
9	2B	H & E: 331, trichrome: 333	Monoclonal antibody directed against LOXL2	PSC
Internal Testing Dataset				
10	2	H & E: 639, trichrome: 633	Insulin sensitizer	Definite NASH; NAS≥4 w/ 1 per component; Fibrosis F1, F2, F3
Held-out Test Set (Standalone Analytical Verification)				
11	2	H & E: 530, trichrome: 532	GLP-1 agonist	Histologic evidence of NASH; Fibrosis F1, F2, F3
12	2	H & E: 900, trichrome: 900	ACC inhibitor, FXR agonist, ASK1 inhibitor	NASH; diagnosis Fibrosis F3, F4

General Model Development Process

The model development followed an iterative process as is illustrated in Figure 11. The development process involved model training, generation of model outputs, and qualitative internal review of outputs. Once satisfactory performance on training data was achieved, the models were deployed on the Internal Test Set (not used in model training) and predefined acceptance criteria were assessed. After meeting the acceptance criteria on the Internal Test Set, models were deployed on the Held-out Test Set and predefined acceptance criteria assessed. The models met the acceptance criteria on the Held-out Test Set and the predefined acceptance criteria were met. The model pipeline was then considered to be locked and validation proceeded.

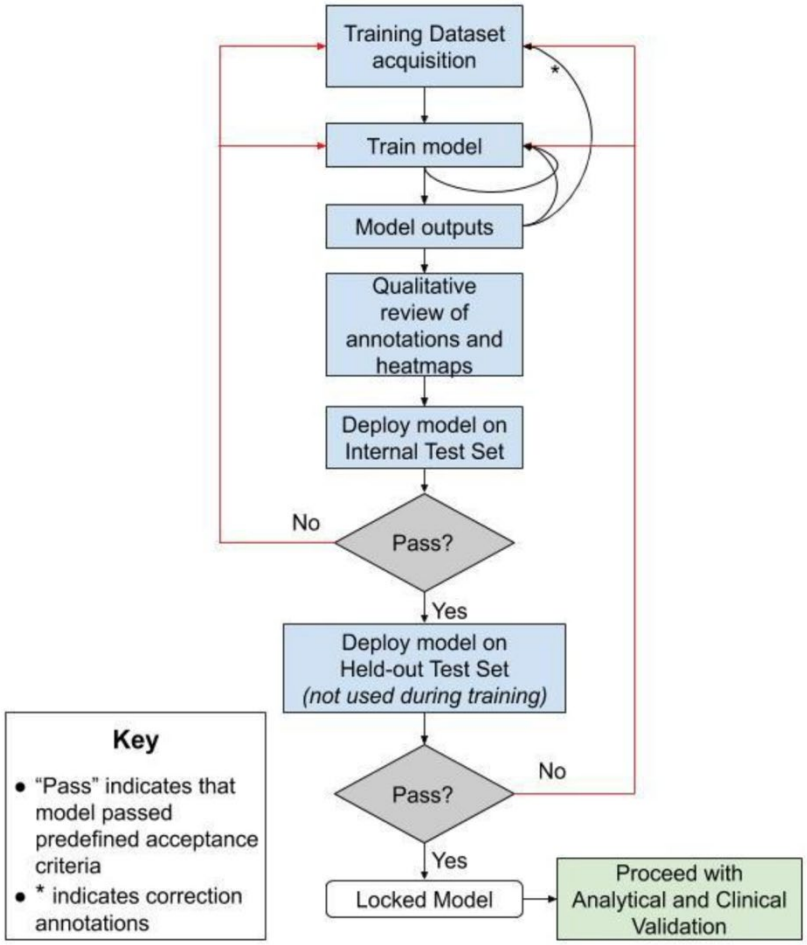
The output of the development process is the list of trained models. These models are described by an architecture, which is stored in a YAML file that lives inside the S3 asset directory, along with the weights file. The architecture and weights together are uniquely specified by the model identifiers (See Table 15)**.**

Referring to these model identifiers is sufficient to allow the model to be applied to new data which is uploaded through to the platform.

For software used in the development process, versions of software used are pinned, and image is frozen. In that way, one can perfectly reproduce the environment if needed to confirm that the same inputs provide the same outputs.

The following sections describe in detail the development of the models comprising the AIM-NASH model pipeline.

Figure 11: AIM-NASH Iterative Model Development



3.7.3 Development of H&E and Trichrome Image Segmentation Models

ML-based image segmentation models were developed for identifying image artifacts and key histologic features of NASH on H&E and trichrome WSIs. Detailed characteristics of the datasets for developing the H&E and trichrome models are provided in

Table 11 and Table 12. Additional slides were used for model development (including liver architectural features and to increase specificity in identifying NASH specific features) from NASH, HBV and PSC Trials for which NAS scores were not available.

Table 11: Dataset Characteristics for H&E Image Segmentation Model Development (population characteristics based on central reader for clinical trials 1-6)

Feature	Training Dataset
Number of Images, n	6227
NAFLD activity score	
NAS<4, n (%)	1140 (18.3%)
N/A	1519 (24.4%)
Steatosis	
0, n (%)	649 (10.4%)
1, n (%)	3738 (60.0%)
2, n (%)	325 (5.2%)
3, n (%)	10 (0.2%)
N/A	1505 (24.2%)
Lobular inflammation	
0, n (%)	57 (0.9%)
1, n (%)	862 (13.8%)
2, n (%)	1842 (29.6%)
3, n (%)	1947 (31.3%)
N/A	1519 (24.4%)
Ballooning	
0, n (%)	838 (13.5%)
1, n (%)	1024 (16.4%)
2, n (%)	2846 (45.7%)
N/A	1519 (24.4%)

Table 12: Dataset Characteristics for trichrome Image Segmentation Model Development (population characteristics based on central reader for clinical trials 1-6)

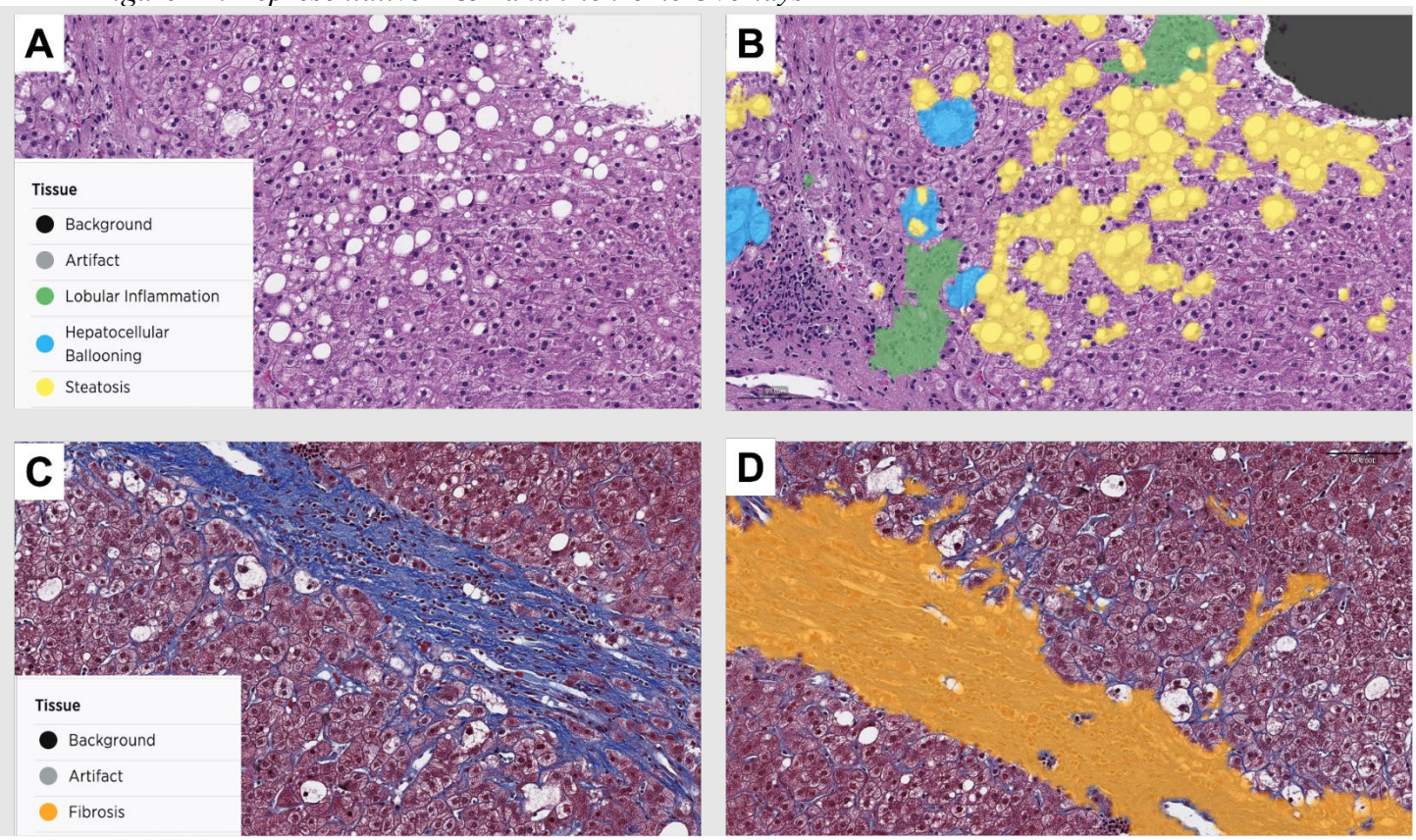
Fibrosis Stage	Training Dataset
Number of Images, n	6215
0, n (%)	222 (3.6%)
1, n (%)	279 (4.5%)
2, n (%)	481 (7.7%)
3, n (%)	1453 (23.4%)
4, n (%)	2173 (35.0%)
N/A	1607 (25.9%)

Table 13 describes the main feature annotations collected and instructions provided to pathologists. Collecting annotations from a large panel of expert pathologists prevented the models from overfitting to a single pathologist's interpretation of the histology. Deep convolutional neural networks were trained, using the annotations generated by our expert pathologists, to identify histological features of NASH. These models were then deployed to produce overlays where each pixel is identified as a specific structure. Annotators were trained to annotate histologic features on H&E and trichrome WSIs using the AISight Translational platform. During

initial rounds for each contributor, roughly 10% of annotations were randomly selected and reviewed for quality by PathAI pathologists. If an annotator had a large number of poor-quality (as defined by incorrect identification of substances by internal expert pathologists) annotations for a particular substance, their annotations for that substance were removed from the dataset.

Specifically, on H&E WSIs, image segmentation models were trained to detect steatosis, lobular inflammation, and hepatocellular ballooning, as well as other hepatic architectural and histologic features (e.g., portal inflammation, micro-vesicular steatosis, normal and pathological trichrome staining, which were identified to increase the robustness of the model). The models generated overlays denoting the presence of the detected histologic features on the WSIs (Figure 12 A and B). On trichrome age segmentation models were trained to detect fibrosis subtypes, including large septal, portal area, subcapsular fibrosis and fibrotic septae (Figure 12 C and D). The models detected and classified the fibrosis subtypes into "pathological" and "non-pathological" fibrosis categories and generated overlays indicating the presence of fibrosis on the WSIs. Non-pathological fibrosis is excluded from fibrosis staging. All overlays generated on training data were assessed for quality by PathAI pathologists, and additional "correction" annotations were collected if necessary, during the iterative model training process.

Figure 12: Representative H&E and trichrome Overlays



Additionally, for quality control, image segmentation models were developed for segmenting artifacts on H&E and trichrome WSIs. Artifacts detected by the models include out-of-focus regions, tissue folds and blur. Finally, areas of artifact are masked from the tissue segmentation models mentioned above.

Detailed Description of Models 1-3: Artifact, H&E Tissue Region Detection, and Trichrome Tissue Region Detection Models

Tissue Region Detection Models were developed for identifying image artifacts (Model 1) and key histologic features of NASH on H&E (Model 2) and trichrome WSIs (Model 3). The Tissue Region Detection Models were required to detect the relevant histologic features prior to scoring using the NASH CRN system. This section will describe in the detail the training dataset characteristics, model training methods and outputs for Models 1-3.

Inputs for Model Training: Models 1-3

The inputs to Models 1-3 were H&E and trichrome WSIs and pathologist annotations of the relevant histologic features. Image-level characteristics of the development dataset are provided in Tables 7 and 8. The development dataset comprised a wide range of NASH severity relevant for trials enrolling both cirrhotic (CRN fibrosis stage 4) and non-cirrhotic (CRN fibrosis stage 1-3) subjects.

Pathologist annotations on H&E and trichrome WSIs were collected from board-certified pathologists specializing in hepatobiliary pathology from the PathAI network of expert pathologists. All pathologists participating in any NASH studies at PathAI are required to meet the following selection criteria:

- Board certification in pathology as evidenced by documentation of ABMS Certification
- Liver pathology subspecialty as evidenced by liver pathology fellowship training and/or significant ongoing clinical experience

In addition, all pathologists providing annotations are required to review the following before contributing data:

- Histologic feature annotations (multiple annotations on one slide)
- NAS grades and CRN fibrosis staging (slide level scores)

Each time a pathologist is assigned a set of cases for review, instructions for completing the review task are provided. These instructions detail how many slides will be reviewed, the type of annotations needed and how many annotations to label per slide.

Region annotations were collected iteratively on training slides using a digital platform (PathAI, Boston, MA). Each pathologist was asked to annotate specific histologic features (e.g., “fibrosis,” “lobular inflammation”) on a set of WSIs. Polygons were created to annotate areas comprising the histologic features of interest. Pathologists were not given specific areas of the WSI to annotate. At no point throughout this process could annotators see any prior annotations on these slides. In training, each annotation was considered an independent training example. If different pathologists happened to annotate the same area independently (even with different labels), these annotations were still considered independent training examples. Note that annotations of histologic features not assessed in the NASH CRN scoring system were included in model training (e.g., portal inflammation, interface hepatitis). Such features enabled the model to differentiate these from features relevant to NASH and fibrosis scoring. A total of 116,346 annotations collected from a panel of 76 board-certified pathologists and used for model training. A total of 58,980 annotations were collected for H&E tissue model (Model 2), 19,075 from

trichrome tissue model and 38,291 from trichrome large septae model (Model 3a and 3b). The pathologists that contributed annotations for model development were trained and qualified according to fit-for-purpose tool development.

Table 13 describes the feature annotations collected and instructions provided to pathologists. In addition to annotations listed in Table 13, pathologists also provided annotations for artifacts (e.g., blur artifact, bubble artifact, ink artifact).

Table 13: Example of Relevant Histologic Feature Annotations Collected by Pathologists

Slide Type	Histologic Feature	Specific Instructions
H&E	NAS: Lobular Inflammation	At least 3 inflammatory cells not including those within sinusoids; Do not label regions of portal inflammation with this region label or regions of interface hepatitis
	NAS: Hepatocellular Ballooning	Please use this label on regions of hepatocellular ballooning. Hepatocellular ballooning is defined as round cells with rarified cytoplasm that are at least 50% larger than neighboring normal cells
	NAS: Macrovesicular Steatosis	Please use this label on regions of dense steatosis. Exclude microvesicular steatosis and any mimickers of steatosis (e.g., glycogenosis, ballooning, swelling)
	Microvesicular Steatosis	Microvesicular steatosis is defined as the presence of nonzonal, contiguous patches of "foamy hepatocytes with centrally placed nuclei". This feature helped the model to differentiate microvesicular from macrovesicular steatosis used for the NAS
	Portal inflammation	Portal inflammation is defined as chronic inflammatory cells restricted to the portal areas. This feature helped the model to differentiate foci of lobular inflammation from inflammation within portal areas.
	Interface hepatitis	Interface hepatitis is defined as inflammatory cells traversing the limiting plate into the adjacent hepatic lobule This feature helped the model to differentiate foci of lobular inflammation from inflammation extending from the portal tract into the periportal zone.
	Normal hepatocytes	Non-steatotic, non-ballooned hepatocytes.
Trichrome	Pathological thickened fibrous septae (Thick fibrotic septae)	Thickened fibrotic septae extending from portal and central regions considered when staging liver biopsies.

Slide Type	Histologic Feature	Specific Instructions
	Pathological fibrosis (portal and perisinusoidal fibrosis)	Portal regions expanded by inflammation, fibrosis, bile ductule proliferation or any combination of the above. Perisinusoidal fibrosis, defined as fibrous deposition distributed in the perisinusoidal space spreading from the centrilobular zone towards the portal area.
	Normal portal areas, small	Normal appearing small/medium sized portal regions, not expanded by fibrosis or inflammation
	Large normal septae (Large septae)	Larger intrahepatic normal septae (usually containing larger arteries, veins and bile ducts) that one would not include when staging liver biopsies
	Subcapsular fibrosis	Normal subcapsular regions of fibrosis not considered when staging liver biopsies

Note that the definitions listed in Table 13 served as a guide for pathologists to annotate the corresponding histologic features of interest at the pixel-level and helped to ensure that the annotations collected from the group of pathologists were consistent and of highly quality. The definitions for the histologic features were determined after consultation with expert hepatobiliary pathologists and review of the relevant literature. Since pathologists were asked to identify and annotate features at the pixel level for the primary annotation step, the guidance in Table 13 was different than that provided in the NASH CRN scoring guidelines which is relevant to scoring at the slide-level. Additionally, the NASH CRN system does not provide definitions of the histologic features themselves but instead is a system from mapping the slide-level prevalence of the histologic features to an ordinal score. Slide-level CRN scores according to the scoring guidelines were collected as a part of the 2nd layer of models (Figure 9), employing graph neural networks, as described further below (Models 4 and 5).

Model Architecture, Training, and Inference

Annotations were grouped into classes as appropriate and then used to generate training sets of image patches on the order of 500,000 samples. These patches were used to train a deep CNN with stochastic minibatch gradient descent using the ADAM optimizer (30) to produce pixel level predictions of NAS components (steatosis, lobular inflammation, and hepatocellular ballooning, fibrosis). Models 1-3 are applied at test time in a patch-wise manner for classifying each pixel within the WSI via a sliding-window approach. No aggregation is applied as each pixel is classified independently. Details of how the trained models are applied to a testing image are provided in

Table 14.

Table 14: Details of Applying Trained Models to a Testing Image

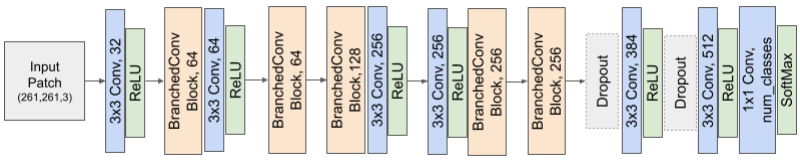
Model Aspects	Model 1 (Artifact Model)	Model 2 (H&E Tissue Model)	Model 3a (Trichrome Tissue Model)	Model 3b (Trichrome Large Septae Model)
Inference Patch-size (Model input size)	261x261	526x526	526x526	526x526
Inference/ Training patch resolution (microns per pixel)	1.0	0.5	0.5	2
Stride	2	8	16	8
Output heatmap resolution (microns per pixel)	2	4	8	16

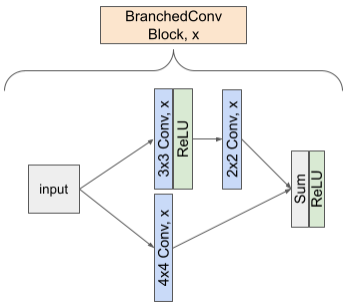
Models are comprised of 8-12 blocks of compound layers with a topology inspired by residual networks and inception networks with a softmax loss (31). While training the models, a set of data normalization and data augmentation steps including zero-mean normalization, random crops, random flips, rotations, HSV transformations and random noise corruption were performed to increase the variance of the data which in turn improved model generalization. A pipeline of image augmentations was used for all segmentation models (1-3). For each training patch, augmentations were uniformly sampled from the following options and applied to the input patch, forming training examples. The augmentations included random crops (within padding of 5 pixels), random rotation (<= 360 degrees), color perturbations (hue, saturation, and brightness), and random noise addition (Gaussian, binary-uniform). After application of augmentations, images were zero-mean normalized. Specifically, zero-mean normalization is applied to the color channels of the image transforming the input RGB image with range [0-255] to BGR with range [-128-127]. This transformation is a fixed reordering of the channels and subtraction of a constant (-128) and requires no parameters to be estimated. This normalization is applied identically to training and test images.

Graphical illustrations of Models 1-3 are provided in the

Figure 13 and Figure 14 below.

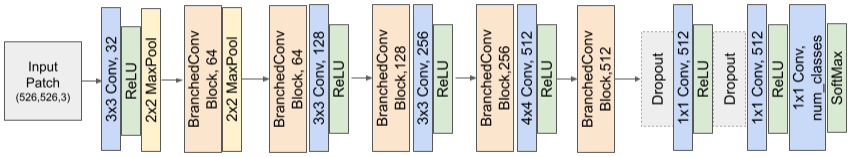
Figure 13: Model 1 (Artifact Detection) Graphical Illustration





Conv: convolutional layer
ReLU: rectified linear unit
"3x3 Conv, 32" indicates a convolutional layer with kernel size 3x3 and 32 output channels (number of filters)

Figure 14: Models 2, 3a, 3b (Tissue segmentation) Graphical Illustration



Note: The same "BranchedConv Block" from Model 1 is used here as well.

The development data was partitioned (labeled Training Datasets in Table 9) into a training set and an internal validation set where the training set was used to fit model weights and the validation set was used as described for monitoring training runs and determining hyperparameters. This partition was made by randomly splitting the training data while balancing for relevant metadata including dataset source and NAS component and fibrosis scores supplied by pathologist reads to ensure each attribute was similarly distributed.

Model 1 was initialized using weights from another model (transfer learning). The initialization model for Model 1 was a previously trained artifact detection developed in other disease contexts. Models 2-3 were trained from scratch using a uniform weight initialization scheme inversely proportional to the input channels and kernel size.

List of hyperparameters include learning rate, batch size, epochs/iterations, optimizer parameters, learning rate scheduler parameters, severity of data augmentation and regularization parameters are tuned using model performance on validation sets for each model. The specific values of these hyperparameters in the final trained algorithm are in Table 15.

Table 15: Hyperparameters of Models 1-3

Category	Hyperparameter	Model 1 (Artifact Model)	Model 2 (H&E Tissue Model)	Model 3a (Trichrome Tissue Model)	Model 3b (Trichrome Large Septae Model)
Learning Rate Parameters	Base Learning Rate	0.001	0.0001	0.0001	0.001
	Learning Rate Scheduler	Staircase	Staircase	Staircase	Staircase
	Learning Rate Decay Factor	0.5	0.5	0.5	0.5
	Learning Rate Decay Steps	2500	10000	10000	5000
Batch Size	Train Batch Size	34	100	100	42
Batch Norm	Momentum Value	0.6	0.6	0.6	0.6
Optimizer	Optimizer Name	Adam	Adam	Adam	Adam
	Optimizer Epsilon	1e-4	1e-4	1e-4	1e-4
Dropout	Dropout Probability value	0.5	0.5	0.5	0.5

On H&E WSIs, tissue region detection models were trained to detect steatosis, lobular inflammation, and hepatocellular ballooning. On trichrome WSIs, tissue region detection models were trained to detect fibrosis subtypes, including large septal, portal area, subcapsular fibrosis and fibrotic septae. The models detected and classified the fibrosis subtypes into "pathological" and "non-pathological" fibrosis categories.

Cloud-computing infrastructure allowed massively parallel patch-wise inference to be efficiently performed exhaustively on every tissue-containing region of a WSI, with a spatial precision of 4-8 pixels. The resulting "overlays" represent model predictions at each point in the WSI.

Model Outputs

The outputs of the tissue region detection models were overlays of steatosis, lobular inflammation, and ballooning on H&E WSIs, and fibrosis on trichrome WSIs. As described above, the overlays are pixel-level model predictions (for plan to validate the overlays, see Section 4.6- Overlay Validation). Overlays generated on training data were assessed for quality by PathAI pathologists, and additional "correction" annotations were collected if necessary, during the iterative model training process (illustrated in Figure 11). Representative overlays on H&E and trichrome WSIs are shown in Figure 12.

3.7.4 Development of Models to Provide NAS Score Components and CRN Fibrosis Scores

Once the overlays are generated for H&E and trichrome-stained slides, the scoring models were trained using graph neural networks (GNN), with NAS component and CRN fibrosis scores from 10 expert liver pathologists as input. As a result, the spatial distributions of detected histologic features were mapped to ML-derived CRN scores for steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis. The NASH scoring models employing GNN were trained using a subset of the training dataset for the image segmentation models. Dataset characteristics are provided in Table 16 and Table 17.

Table 16: Characteristics of NAS Scoring (GNN) Model Development Datasets

Feature	**Training Dataset**	**Internal Testing Dataset (consensus)**	**Total**
Number of Images, n	1530	639	2169
Steatosis			
0, n (%)	132 (8.6%)	25 (3.9%)	157 (7.2%)
1, n (%)	724 (47.3%)	198 (31.0%)	922 (42.5%)
2, n (%)	465 (30.4%)	222 (34.7%)	687 (31.7%)
3, n (%)	209 (13.7%)	187 (29.3%)	396 (18.3%)
N/A	0 (0%)	7 (1.1%)	7 (0.3%)
Lobular inflammation			
0, n (%)	205 (13.4%)	20 (3.1%)	225 (10.4%)
1, n (%)	879 (57.5%)	373 (58.4%)	1252 (57.7%)
2, n (%)	369 (24.1%)	231 (36.2%)	600 (27.7%)
3, n (%)	77 (5.0%)	8 (1.3%)	85 (3.9%)
N/A	0 (0%)	7 (1.2%)	7 (0.3%)
Hepatocellular ballooning			
0, n (%)	417 (27.3%)	87 (13.6%)	504 (23.2%)
1, n (%)	613 (40.1%)	276 (43.2%)	889 (41.0.%)
2, n (%)	500 (32.7%)	268 (41.9%)	768 (35.4%)
N/A	0 (0%)	8 (1.2%)	8 (0.4%)
NAFLD activity score			
NAS<4, n (%)	649 (42.4%)	148 (23.2%)	797 (36.8%)
NAS≥4, n (%)	881 (57.6%)	483 (75.6%)	1364 (62.9%)
N/A	0 (0%)	8 (1.2%)	8 (0.4%)

Table 17: Characteristics of Fibrosis Staging Model Development Datasets

Fibrosis stage	Training Dataset (image n=1292)	Internal Testing Dataset (consensus) (image n=633)	Total (image n=1925)
0, n (%)	59 (4.6%)	15 (2.4%)	74 (3.8%)
1, n (%)	172 (13.3%)	159 (25.1%)	331 (17.2%)
2, n (%)	186 (14.4%)	146 (23.1%)	332 (17.3%)
3, n (%)	483 (37.4%)	278 (43.9%)	761 (39.5%)
4, n (%)	392 (30.3%)	23 (3.6%)	415 (21.6%)
N/A	0 (0%)	12 (1.9%)	12 (0.6%)

3.7.5 Detailed Description of Models 4 and 5: NASH CRN Scoring Models

ML CRN scoring models were developed for scoring the NAS components from H&E WSIs (Model 4) and fibrosis on trichrome WSIs (Model 5). These models used the histologic features detected by the tissue region detection models (Models 1-3) as inputs to generate image-level NASH CRN component scores, which are the primary read-outs of AIM-NASH. This section will describe in detail the training dataset characteristics, model training methods and outputs for the NASH CRN scoring models.

Inputs for Model Training

The data used for training the NASH CRN scoring models were the overlays generated on H&E and trichrome WSIs from Models 1-3 and corresponding NASH CRN component scores provided by NASH pathologists from the PathAI expert contributor network. NASH CRN scores were collected from 10 trained pathologists. The process of collecting NAFLD slide-level scores used the PathAI digital platform. Pathologists were presented with the WSI alone and were unable to view any model outputs or scores from other annotators. For training of Models 4-5, each WSI graph, derived from Models 1-3 outputs, and pathologist slide-level component label was considered a training example to the GNN models. In other words, the graph is the input to the GNN and the slide-level component label is the prediction target. For slides where there were two or more independent scores for the same component, each score and associated graph was used as an independent training example.

The reason for leveraging models instead of annotations during graph construction is that graph construction for the GNNs requires exhaustive labeling of the entire slide's tissue and exhaustive annotation by human pathologists is time consuming and difficult to reproduce or perform accurately. In addition, during inference the GNN models predict NAS and CRN scores based on input graphs derived from CNN model outputs, and therefore this approach is mimicked during training. Pixel annotations are only used for training Models 1-3. Models 4-5 are exclusively trained on disease severity scores (NASH CRN scores) provided by human pathologists. WSI pixels were clustered to form superpixels using Birch clustering after filtering using the outputs from Models 1-3. Specifically, overlays are 3 dimensional matrices with dimensions corresponding to X-coordinate, Y-coordinate, and CNN model outputs for each pixel in the WSI. Pixels were filtered such that clusters are formed

only of pixels classified as a NASH relevant class (i.e. pixels classified as artifact, background, or usable tissue are ignored.) The remaining pixels were clustered using the Birch algorithm in 2D using X and Y coordinates. These resulting clusters or "superpixels" are the nodes of the graph.

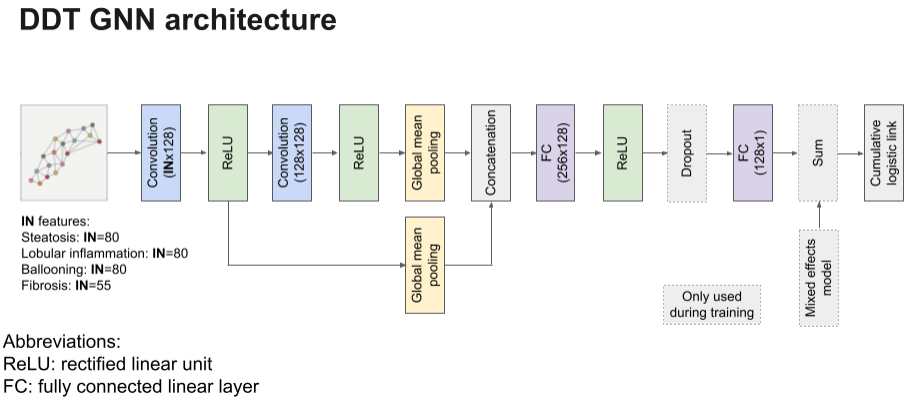
Model Architecture, Training and Inference

Models 4-5 utilized Graph Neural Networks (GNNs) to predict slide-level NASH CRN scores. GNNs are an emerging deep learning method that represent and characterize histologic features using graph representations and are well-suited to data types that can be modeled by a graph structure, such as fibrosis architecture (32). The H&E and trichrome WSI overlays of relevant histologic features, produced by the CNN as described above, were clustered into "superpixels" to construct the nodes in the graph, reducing hundreds of thousands of pixel-level predictions into thousands of super-pixel clusters. WSI regions predicted as background, artifacts or normal tissue were excluded during clustering. Directed edges were placed between each node and their 5 nearest neighboring nodes (via the K Nearest Neighbor algorithm) with a maximum distance cutoff of 0.25 x minimum (image height, image width) to prevent edges being placed between clusters from different tissue slices. Each graph node is represented by three classes of features generated from previously trained CNN predictions pre-defined to be biological classes of known clinical relevance. Spatial features included the mean and standard deviation of (x, y) coordinates. Topological features included area, perimeter, convexity of the cluster. Logit-related features included the mean and standard deviation of logits for each of the classes of CNN generated heatmaps. Following construction, each WSI was represented by a single directed graph. The GNN performed a graph-level ordinal classification of CRN NAS component scores (from H&E WSIs) and CRN fibrosis stage from (trichrome WSIs).

In addition, GNNs in Models 4-5 learned the individual pathologist's performance parameters and corrected any biases during the deployment to resolve the pathologist discordance in scoring NASH CRN scores. More specifically, the GNN used a "mixed effects" model where each pathologist's bias was specified by a set of parameters learned during training. The model generates a score by selecting the specified pathologist bias parameter and adding it to the unbiased estimate of the NAS component. Upon deployment, the labels are produced using only the unbiased estimate.

A graphical illustration of Models 4-5 is provided in Figure 15.

Figure 15:Models 4a-c and 5 (GNN Scoring)



Data was split into training, internal validation and test set prior to training the GNN (same splits used for both CNN and GNN) where the NASH CRN scores are balanced across these splits. The optimal hyperparameters are chosen based on the performance of the trained model on the validation data set and were completely agnostic to the test set.

Models 4-5 were initialized using He-weight initialization. The list of hyperparameters included learning rate, batch size, epochs/iterations, optimizer parameters, learning rate scheduler parameters and regularization parameters. The specific values of these hyperparameters in the final trained algorithm are in Table 18.

GNN specific hyperparameters included graph construction hyperparameters (e.g. number of nodes in the graph) as well as additional model training hyperparameters (e.g. graph convolution parameter, graph pooling parameter, correction of pathologist bias).

Table 18: Hyperparameters of Models 4-5

Parameter	**Hyperparameter**	**Model 4a Steatosis**	**Model 4b Ballooning**	**Model 4c Lobular inflammation**	**Model 5 Fibrosis**
Learning Rate Parameters	Base Learning Rate	0.001	0.001	0.001	0.001
	Learning Rate Scheduler	Staircase	Staircase	Staircase	Staircase
	Learning Rate Decay Factor	0.8	0.8	0.8	0.8
	Learning Rate Decay Steps	200	200	200	200
Batch Size	Train Batch Size	32	32	32	32
Batch Norm	Momentum Value	Adam	Adam	Adam	Adam
Optimizer	Optimizer Name	0.9 / 0.999	0.9 / 0.999	0.9 / 0.999	0.9 / 0.999
	Optimizer Epsilon	1e-8	1e-8	1e-8	1e-8
Dropout	Dropout Probability value	0.5	0.5	0.5	0.5
Network Architecture	Hidden features	128	128	128	128
	Layers	2	2	2	2
Mixed effect model	Bias multiplier	0.1	0.01	0.07	0.1

Model Outputs

The outputs of the NASH CRN scoring models are NASH CRN scores for steatosis, ballooning, and lobular inflammation grade (from H&E WSIs; Model 4) as well as fibrosis stage (from trichrome WSIs; Model 5).

3.7.6 Limitations in Model Development

While the AIM-NASH model was trained using a patient population representative of those enrolled in NASH trials, there are some limitations to address. ML methods are mostly limited by diversity in the training & validation data. The following data types are limited in our dataset:

- Artifacts that were not encountered during development, such as various color markers, unknown scanner artifacts
- Non-liver organ types including skin, smooth muscle, and kidney
- Staining and scanner variations that were not encountered during development
- Very little useable liver tissue available on the slide to score
- Biopsy being broken into multiple fragments

Machine learning algorithms can be understood by defining their internal structure and having clear knowledge about the framework of inputs and the relation to their outputs. The risks of possible over- and underfitting of algorithms need to be acknowledged and balanced with the advantages that well-designed algorithms can provide. PathAI mitigates this risk by using a training, validation, and testing framework.

3.7.7 Change Control

Any changes, including bug fixes or system patches, to locked and validated AIM-NASH are evaluated to determine the need for additional analytical or clinical validation. This evaluation is conducted according to PathAI's Change Management Procedure, which is aligned with relevant regulations and internationally recognized consensus standards. Any proposed changes are evaluated in the context of potential impact, both direct and indirect, and classified as either Major or Minor, with Major changes requiring revalidation. The Revision Level History table for AIM-NASH is included in Appendix VIII.

4 Methodology and results

For clarity in platform names throughout this section and in related appendices, please refer to this table.

Current platform name	**Prior names utilized in EMA communications and supporting documentation**
AISight Clinical Trials	Portal, AISight Clinical Trials Services, AISight CTS
AISight Translational	Slides Platform

4.1 Standalone analytical Verification (SAV)

4.1.1 Objectives and Methodology

SAV of AIM-NASH on the AWS development environment was performed to confirm that the algorithm meets specified acceptance criteria outlined below. The algorithm was deployed on two held-out test datasets, which were not used for model training (Table 19).

Table 19: Held-out datasets used for SAV

Trial Name & Sponsor	Phase	Fibrosis Stage at Trial Entry	Slides Available for testing
ATLAS Gilead	2	F3, F4	900 H&E; 900 trichrome
Semaglutide Novo Nordisk	2	F1, F2, F3	530 H&E: trichrome: 532

A single case is comprised of 1 H&E and 1 trichrome slide. Out of the available data set (Table 22) 250 cases were selected based on fibrosis stage derived from AIM-NASH derived scores, with approximately 50 cases being represented per fibrosis stage (47 for F0, 53 for F2, 50 each for F1, F3, & F4). Of these cases, 250 trichrome slides and 249 H&E slides were available. A reference standard was developed for all NAS components (steatosis, lobular inflammation, and hepatocellular ballooning) and fibrosis stages based on three pathologists providing manual scores on all slides and generating a consensus score and is summarized in Table 21.

Acceptance Criteria

The acceptance criteria required the lower 2.5% confidence interval of the linearly WK of the AIM-NASH scores vs. the reference standard median consensus scores be at least as good as 0.1 below the mean pairwise linearly WK among network pathologists.

1. Accuracy was assessed separately for each NAS component (steatosis, lobular inflammation, and hepatocellular ballooning,) and fibrosis stage.
2. The benchmark scores compared against were the pairwise linearly weighted Kappas generated during SAV from pathologists.

Suspension Criteria

If numerous or severe failures of AIM-NASH were discovered, it may have been necessary to suspend testing, and resume after assignable cause has been determined. The decision to suspend testing and the methodology for resuming must be documented, including a rationale for doing so. The decision for suspending the verification would be employed for failures that would affect PathAI's ability to use AIM-NASH and maintain compliance.

Verification Strategy

The extent of testing, performed by ML engineers, is detailed in the test protocols and is dependent on the risk of failure of the feature being tested *Table 20*. Only once SAV is completed successfully, AIM-NASH can be

integrated onto the AISight Clinical Trials platform. If the results are not acceptable, then SAV would be performed again with a new set of held-out slides.

Table 20: Requirements and Tests

Execution Steps	1. Deploy AIM-NASH to generate NAS component scores on H&E slides and fibrosis stage on trichrome slides. 2. Calculate the agreement between the AIM-NASH and the manual read reference standard.
Expected Results	1. Successful generation of scores for slide set. 2. Agreement meets acceptance criteria specified *above*.

Table 21: Slide distribution for SAV by 3-way consensus

Trial Name & Sponsor	Grade/Stage	Steatosis	Lobular inflammation	Hepatocellular ballooning	Fibrosis
ATLAS Gilead	0	12	6	14	21
	1	51	72	66	32
	2	31	37	36	19
	3	22	1	-	21
	4	-	-	-	21
Semaglutide Novo Nordisk	0	9	4	16	13
	1	51	73	61	37
	2	40	37	38	23
	3	15	1	-	27
	4	-	-	-	6

4.1.2 SAV Results

Consensus scores were generated on 231 H&E slides and 220 trichrome slides, the remaining slides (19 H&E and 29 trichrome slides) were deemed non-evaluable by consensus reads. The algorithm was tested as according to the SAV plan and the activities identified therein were completed and all acceptance criteria were met (Table 22). It can be concluded that the algorithm results generated during SAV on the AWS development environment met the prespecified acceptance criteria (*Linear Kappa Values (CI 2.5, 97.5%), excluding cases deemed Non-Evaluable (NE) by Reference Pathologists)*.

Table 22: Agreement of AIM-NASH consensus readouts and pathologist mean pairwise comparison

NASH Component	N	AIM - Consensus WK	Pathologist Mean Pairwise WK
Steatosis	231	0.68 (0.62, 0.75)	0.55 (0.5, 0.6)
Lobular inflammation	231	0.5 (0.42, 0.58)	0.45 (0.37, 0.51)
Hepatocellular ballooning	231	0.49 (0.41, 0.56)	0.39 (0.32, 0.45)
Fibrosis	220	0.7 (0.65, 0.74)	0.65 (0.62, 0.69)

4.2 Integrated analytical verification (IAV)

4.2.1 Objectives and Methodology

Software verification and IAV of AIM-NASH v1.1.0 on the AISight Clinical Trials Platform (the platform utilized in clinical validation for AI-assisted reads) was performed to confirm that AIM-NASH results are viewable in the expected format, and to confirm that AIM-NASH results generated during SAV on the AWS development platform agree with the results generated on AISight Clinical Trials for the same slide set.

Software verification was conducted on 1 H&E and 1 trichrome slide scanned at 20x, and 1 H&E and 1 trichrome slide scanned at 40x (for this submission, only the 40x scanned image verification is relevant as per the proposed workflow in clinical trials). For IAV 20 trichrome slides and 20 H&E slides from the held-out test set used in SAV were used from Table 19.

Acceptance Criteria

The Locked Model of the AIM-NASH algorithm shall yield the same results on the AISight Clinical Trials platform upon integration as it did on the development environment for the held-out test set.

Suspension Criteria

If numerous or severe failures of AIM-NASH are discovered, it may be necessary to suspend testing, and resume after assignable cause has been determined. The decision to suspend testing and the methodology for resuming must be documented, including a rationale for doing so. The decision for suspending the verification would be employed for failures that would affect PathAI's ability to use the software and maintain compliance.

Verification Strategy

The extent of testing is detailed in the test protocols and is dependent on the risk of failure of the feature being tested.

1. Software Verification: The verification strategy is to evaluate that the results specified in product requirements are visible in the case viewer.
2. IAV: The verification strategy is to determine that the results from SAV performed on the AWS development environment are equivalent to the results of AIM-NASH deployment on the Clinical Trials Platform. This will be achieved by deployment of the AIM-NASH on a subset of the slides used for SAV.

Overall Risk Justification

PathAI has determined that there is a Medium risk that hazards may cause patient harm due to the failure of the integrated AIM-NASH algorithm and viewing platform product, as AIM-NASH is used for enrollment and

follow-up timepoint scoring in NASH clinical trials (Table 23). The extent of verification is based on the risk of patient harm.

Table 23: Risk determination

Risk Severity	**Risk Definition**
High (H)	If a failure associated with a User Requirement occurred, there would be direct impact on patient safety, product quality or the data integrity
Medium (M)	If a failure associated with a User Requirement occurred, there would be an indirect impact on patient safety, product quality or the data integrity.
Low (L)	If a failure associated with a User Requirement occurred, there would be no impact on patient safety, product quality or the data integrity.

Required Tests

Test steps for each of the below verification plans are detailed in the results section 4.2.1.2, Appendix IX, and Appendix X, respectively.

- Software Verification Plan – Slides scanned at 20X
- Software Verification Plan – Slides scanned at 40X
- Integrated Analytical Verification Plan

4.2.2 IAV Results

For AV on the platform, the locked model was deployed on the held-out test subset. The results from this activity must match the results of SAV, for which the same model was deployed on the held-out test set in the development environment. The purpose of IAV was to ensure that the locked model yielded the same results with defined tolerance on the platform as it did in the development environment, using the same held-out test subset to verify that the platform integration, functional, user workflow and reporting requirements defined for the algorithm product are met.

AIM-NASH integration into the AISight Clinical Trials platform was tested successfully in two cycles against the acceptance criteria in the IAV plan. The second cycle was executed as there were issues observed during software verification and IAV. During execution of Test Step ID 64, it was found that one sample returned double results and overlays for both the H&E and trichrome slides on the Slide Viewer screen. Although this was not a test failure, engineering was contacted to determine why this occurred for this one sample. Engineering representative indicated that the algorithm was triggered twice for that sample. A fix for the issue was

implemented and test cycle 2 was executed to verify this fix. Test cycle 2 demonstrated that the fix has been implemented and there is no impact on the tool.

It can be concluded that AIM-NASH met the requirements specified by the product requirements and the acceptance criteria in IAV Plan and Protocol. The AIM-NASH algorithm produced equivalent results for the same 20 slide sets in the AISight Clinical Trials Platform as the AIM-NASH Algorithm v1.1.0 produced during SAV in the ML Platform environment. The conclusion of this verification is that AIM-NASH Algorithm v1.1.0 is acceptable for use as intended.

Nonconformances

The identified nonconformances represent disruptions for time of workflow, and do not greatly impact the safety and efficacy of the tool. Detailed reporting of nonconformances are provided in Appendix IXb.

4.3 Validation of the AISight Clinical Trials Platform

4.3.1 Product Description

The AISight Clinical Trials platform (v3.3.1) is a RUO cloud-based software as a service (SaaS) platform that enables PathAI partners to utilize PathAI artificial intelligence (AI)-powered algorithms in prospective clinical trials, supporting eligibility and stratification, response monitoring, exploratory analysis, and quality control use cases at scale. Platform configurability allows for maximum flexibility in leveraging digital pathology to improve subject outcomes in clinical research. Pathology evaluations performed on the AISight Clinical Trials platform for clinical trials should not be used to inform patient care outside of the clinical trial and are for research purposes only.

The AISight Clinical Trials platform is a secure, web-based platform intended to be used by the following qualified medical and laboratory professionals: sponsors, trial managers, site managers, pathologists, histotechnologists, data managers, and study monitors. Each user can execute role-specific actions to upload, process, and view samples as slide images in deep zoom image (DZI) format. The platform includes tools for generating AI-powered subject-level and lab-level reports that provide a detailed analysis. Some of these reports include quantitative assessments of histological features, including heterogeneity, cell density and spatial relationships.

These assessments enable users to identify subjects, monitor drug activity, and predict subject outcomes more accurately than manual scoring methods, which may be prone to inter and intra-observer variability.

The AISight Clinical Trials platform was utilized in collection of AI-assisted reads for the purposes of this qualification.

4.3.2 Objectives

The primary objective of the AISight Clinical Trials platform validation study was to validate the platform for NASH reads using glass slides scanned on the Aperio AT2 (Leica) whole slide scanner by evaluating non-inferior agreement of NASH (defined as NAS $\geq$ 4 with a score of $\geq$ 1 for each component and absence of atypical features suggestive of non-NASH liver disease, similar to the definition used during NASH clinical trial enrollment) and

non-NASH diagnosis between glass GT read and WSI read versus agreement between glass GT read and individual study pathologist glass read.

4.3.3 Study Design and Plan

PathAI utilized existing de-identified glass slides from a third-party vendor (Precision for Medicine) and from partners from their completed clinical trials (screen failures from Phase 2B study from Northsea Therapeutics NCT04052516 and enrolled cases from a Phase 2 study from Madrigal Pharmaceuticals NCT02912260). Each case utilized in this study had 2 slides per case, including one H&E- and one trichrome-stained slide. Slides were first scanned at the PathAI Biopharma Lab in Memphis on the Aperio AT2 (Leica) whole slide scanner at 40x magnification and then distributed for glass reads.

The platform was validated by 6 board-certified hepatopathologists who had demonstrated proficiency in reading manual NASH cases (see Appendix IIb for more details on proficiency documentation in PathAI's eQMS). The GT reads were collected on glass slides using a microscope by 3 board certified hepatopathologists from PathAI Contributor Network. These 3 pathologists had experience in reading NASH cases in their clinical practice and for PathAI projects. Each of these pathologists read 160 cases on glass once. The glass GT score was computed as the median of all 3 scores. Additionally, the majority of the 3 GT pathologists' responses were used to assess presence of atypical features.

A different set of 3 board certified hepatopathologists performed the study reads. They read all cases twice, once on glass using a microscope and once on WSIs using the platform, with a minimum of 2-week washout between reads with different modalities. See Figure 16 and Figure 17 for study design and logistics.

The slide set consisted of 160 cases (320 slides - each case consists of an H&E slide and a trichome slide) from liver needle biopsies. Two thirds of the cases were chosen from patients with NASH (defined as NAS ≥4 with a score of ≥1 for each component: steatosis, lobular inflammation and hepatocellular ballooning and absence of atypical features suggestive of non-NASH liver disease) based on the original trial central pathology scores, and the remaining one third are from NAFLD (without NASH) and other (non-NAFLD) liver indications, including but not limited to hepatitis B, hepatitis C, active hepatitis, viral hepatitis, cirrhosis and intrahepatic cholestasis. Five to ten percent of the cases were chosen to be challenging, defined as NAS ≥4 with a score of 0 for at least one of the components (steatosis, lobular inflammation, and hepatocellular ballooning), NAS =4 with a score of ≥1 for each of the components (steatosis, lobular inflammation, and hepatocellular ballooning) or NAS =3. For glass reads, the 160 cases were split into 3 batches and the pathologists read 1 batch at a time.

Figure 16: Study Design

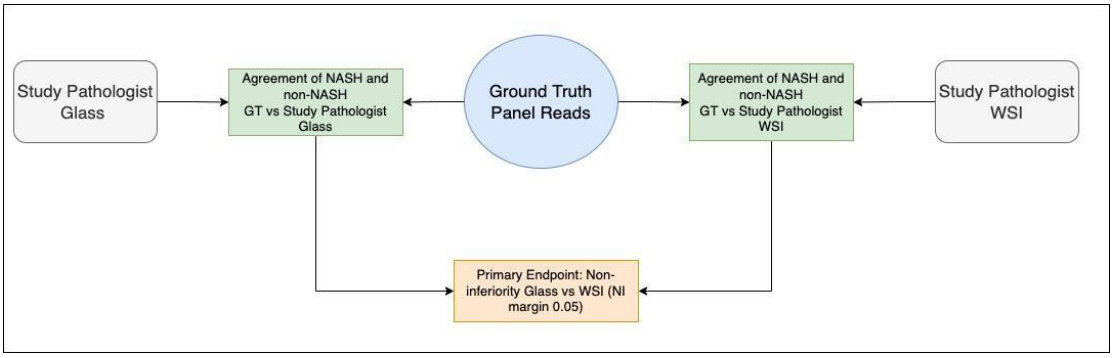
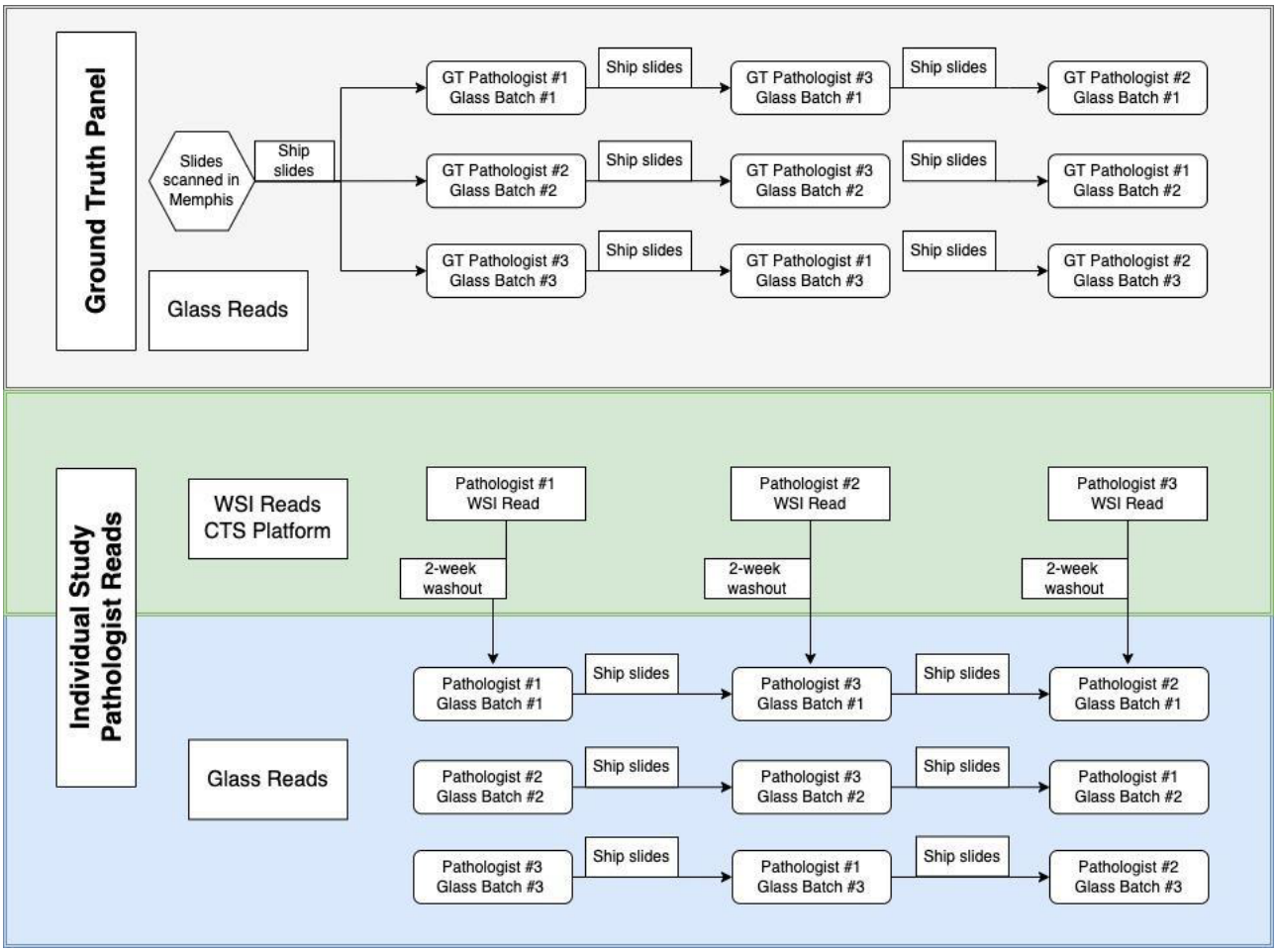


Figure 17: Study Logistics



4.3.4 Dataset

Three hundred and eighteen (318) slides (159 cases) were enrolled in the study by the GT pathologists. The glass slides were from completed Phase 2 studies (enrolled samples from Madrigal's non-cirrhotic phase 2 study for thyroid hormone receptor-b agonist and screen failures from NorthSea Therapeutics non-cirrhotic phase 2b study for carboxylic acid) and liver biopsy glass slides purchased from Precision for Medicine. Overall, 96 slides were enrolled from NorthSea, 186 slides were enrolled from Madrigal and 36 slides were enrolled from Precision for Medicine.

Table 24: Study population

Source/Trial Phase	Drug Class	Enrollment Criteria
Madrigal MGL-3196 Phase 2 enrolled population	thyroid hormone receptor-β agonist	● NAS ≥ 4, with a score of at least 1 in each component (steatosis, lobular inflammation, and ballooning) ● Fibrosis stage 1 to 3
NorthSea Icona Trial, Phase 2B Screen-failed population only	Carboxylic acid	● NAS ≥ 4, with a score of at least 1 in each component (steatosis, lobular inflammation, and ballooning) ● fibrosis score F1 to F3, inclusive (F1 capped at 30%)
Precision for Medicine	N/A	-

4.3.5 Selection of Study Population/ Cases

Inclusion Criteria

- Study slides scanned at pre-specified magnification (40x).
- Images are quality checked according to scanner instructions and can be rescanned if needed. All information logged by the lab technician and captured per lab established SOPs.
- Two (2) slides included per case: one H&E slide and one trichrome slide.
- One case per patient, i.e., unique cases.
- De-identified cases.
- Only liver biopsies included.

Exclusion Criteria

- Cases for which the slides do not fulfill quality check according to the scanning acquisition device.
- Cases with indelible markings.
- Cases with any patient identifying information.
- Cases with any tissue other than liver FFPE tissue.
- Cases without both H&E and trichrome slides available.

4.3.6 General Procedures

Blinding

All participating pathologists had their own unique log-in to the AISight Clinical Trials platform and OpenClinica electronic data capture (eDC) platform and were assigned their specific study cases. The pathologists were blinded to each other's assessments and to their own assessments from different modalities (glass and digital reads). All PathAI staff (except for the unblinded clinical data managers and unblinded clinical scientist) involved in this study were blinded to the data until the database was locked. No PHI data was collected in this study.

Glass Slides Scanning and Handling

Glass slides utilized for this validation study were from completed Phase 2 studies (enrolled samples from Madrigal Pharmaceuticals MGL-3196 study NCT02912260, screen failures from NorthSea Therapeutics ICONA

clinical study NCT04052516) and glass slides purchased from Precision for Medicine, a commercial biobank that collects their samples under IRB oversight following the highest industry standards. These slides were labeled and scanned by PathAI Biopharma lab on an Aperio AT2 whole slide scanner at 40X magnification. The Biopharma Lab was also responsible for shipping the glass slides to the 6 participating pathologists and generating shipping labels for each shipment. The lab also re-labelled the glass slides with the original trial information after the completion of the study.

4.3.7 Pathologist Training

All pathologists were trained on the study protocol and required tasks (See Appendix Vc for Study Case Report Forms) by the Principal Investigator (PI) prior to participating in any study activities. All pathologists also signed an attestation form acknowledging the completion of training. All training records are stored in PathAI's electronic Quality Management System (eQMS).

4.3.8 Data Handling

All data was entered electronically in the AISight Clinical Trials platform for the manual digital reads and OpenClinica for the glass reads. After the completion of the study, all data was securely downloaded from the platforms and stored in the clinical data management S3 bucket. Data for analysis was uploaded to PathAI's eQMS once the database lock form was approved. PathAI's designated unblinded clinical data manager and clinical scientist had access to all study information for the purpose of monitoring data and resolving any queries.

Data was downloaded from the database service for the AISight Clinical Trials platform and OpenClinica over the course of the study for data monitoring purposes. All relevant study data along with corresponding documentation was uploaded to the Clinical Data Management Amazon Web Services (AWS) bucket which only unblinded clinical data managers have access to.

4.3.9 Statistical Methods and Determination of Sample Size

Primary Endpoint

The primary endpoint is the agreement of NASH (defined as NAS $\geq$ 4 with a score of $\geq$ 1 for each component and absence of atypical features suggestive of non-NASH liver disease, similar to the definition used during NASH clinical trial enrollment) and non-NASH diagnosis between glass GT read and WSI compared to the agreement of NASH and non-NASH diagnosis between glass GT read and individual study pathologist glass read.

The null hypothesis is that the agreement of NASH and non-NASH diagnosis between glass GT read and WSI is inferior to the agreement of NASH and non-NASH diagnosis between glass GT read and individual study pathologist glass read discounted by a non-inferiority margin of 0.05. The alternative hypothesis is that the agreement between glass GT read and WSI is non-inferior to the agreement between glass GT read and individual study pathologist glass reads discounted by a non-inferiority margin of 0.05. These hypotheses are stated as follows:

$$H_0: \pi_{GD} \leq \pi_{GG} - 0.05$$

$$H_a: \pi_{GD} > \pi_{GG} - 0.05$$

where π_{GD} is the average agreement across 3 pathologists of NASH and non-NASH diagnosis between glass GT read and WSI and π_{GG} is the average agreement across 3 pathologists of NASH and non-NASH diagnosis between glass GT read and individual study pathologist glass reads. π_{GD} will be shown to be statistically non-inferior to 0.05 less than π_{GG} if Bootstrap percentile $p < 0.025$. Analysis was done on observed data.

Secondary: This endpoint consists of study pathologist scores for the four primary NASH components (NAS components on H&E and CRN fibrosis on trichrome slides), and the overall NAS score between WSI and glass read. This endpoint was evaluated as described below:

Linearly WK concordance statistics between glass and WSI read for each of the pathologists, each of the NASH components, and overall NAS score. Overall, linearly WK was computed for each NASH component and overall NAS score will be computed by averaging the WK for the 3 pathologists. Bootstrap 95% confidence intervals are provided on the overall as well as per pathologist linearly WK. These concordance estimates are compared to the published range in Table 25.

These analyses are based on observed data.

Table 25: WK scores for intra reader variability

Feature	**Publication**	**Intra-observer variability** (WK scores)
Steatosis	Kleiner et al. 2005 (8)	0.83
	Gawrieh et al. 2011 (25)	0.72 (pre)* and 0.75 (post)*
	Davison et al. 2020 (14)	0.666
Lobular inflammation	Kleiner et al. 2005 (8)	0.60
	Gawrieh et al. 2011 (25)	0.37 (pre)* and 0.48 (post)*
	Davison et al. 2020 (14)	0.227
Hepatocellular ballooning	Kleiner et al. 2005 (8)	0.66
	Gawrieh et al. 2011 (25)	0.32 (pre)* and 0.56 (post)*
	Davison et al. 2020 (14)	0.487
Fibrosis	Kleiner et al. 2005 (8)	0.85
	Gawrieh et al. 2011 (25)	0.64 (pre)* and 0.75 (post)*
	Davison et al. 2020 (14)	0.679
NAS	Davison et al. 2020 (14)	0.372

*Pathologists in this study read slides before an intervention and after an intervention. The intervention consisted of a review of illustrative histologic images of NAFLD with the study pathologists and use of scoring sheet with written diagnostic criteria for different NAFLD phenotypes.

Determination of Sample Size

The College of American Pathologists (CAP) guidelines recommend a minimum of 60 cases to ensure that diagnostic performance based on digitized slides is at least equivalent to that of glass slides and light microscopy and to identify and rectify risks associated with the technology (29). With substantial inter-rater variability in NASH scoring and diagnosis, a non-inferiority design was determined to be more appropriate for the NASH trial population than a direct comparison of agreement between glass and digital reads. A sample size of 160 slides was selected to provide a higher degree of precision around the estimates and to account for not evaluable slides, and any incidental breakage of glass slides.

4.3.10 AISight Clinical Trials Platform Validation Results

Study Patients/ Subjects

One hundred and fifty-nine (159) cases were enrolled in the study by the GT pathologists. The glass slides were from completed Phase 2 studies (enrolled samples from Madrigal Pharmaceuticals MGL-3196 study NCT02912260 and screen failures from NorthSea Therapeutics ICONA study NCT04052516) and liver biopsy glass slides purchased from Precision for Medicine.

Demographic and Other Baseline Data

No demographic information for the slides enrolled in the study is available. The dataset represents both failed and enrolled NASH clinical trial patient populations, liver biopsies from other liver diseases (including but not limited to hepatitis B, hepatitis C, active hepatitis, viral hepatitis, cirrhosis and intrahepatic cholestasis) and normal liver. The dataset also contains variability in sample staining (including performed by multiple collection/preparation sites). Distribution of slides based on slide level score from glass GT are listed in Table 26.

Table 26: Distribution of Slides Based on Glass GT

Feature	Score	% (n/N)
Steatosis	0	8.2 (13/159)
	1	32.7 (52/159)
	2	30.2 (48/159)
	3	28.9 (46/159)
Lobular inflammation	0	1.9 (3/159)
	1	62.3 (99/159)
	2	34.0 (54/159)
	3	1.9 (3/159)
Hepatocellular ballooning	0	22.6 (36/159)
	1	56.6 (90/159)
	2	20.8 (33/159)
Fibrosis	0	6.9 (11/159)
	0.5*	0.6 (1/159)
	1	27.7 (44/159)
	2	28.9 (46/159)
	3	28.3 (45/159)
	4	7.5 (12/159)

* Fibrosis stage is 0.5 because median stages for all 3 GT pathologists were used. See section 9.5 of attached Protocol in Appendix Va for further information.

Primary Analysis

One hundred and fifty-nine (159) cases were enrolled in the study by 3 GT pathologists by reading glass slides using a microscope. A separate set of 3 study pathologists evaluated the same 159 cases once as glass slides using a microscope, then as WSIs on the AISight Clinical Trials platform, with a minimum of 2-week washout between the 2 different modalities.

The acceptance criteria for non-inferiority (with a margin of 0.05) agreement for NASH diagnosis between reads on WSI and glass GT compared to reads on glass and glass GT was met with a difference of -0.001 (95% CI, -0.027, 0.026; $p<0.0001$; Table 27). The agreement between study pathologists reads on AISight Clinical Trials

platform using WSIs and glass GT was 0.743 (95% CI, 0.7, 0.788) and the agreement for glass reads and glass GT was 0.745 (95% CI, 0.703, 0.786).

Table 27: Agreement between reads on WSI and glass GT vs reads on glass and glass GT

Comparison	N	Agreement Rate (95% CI)	Difference (95% CI)	P-value
WSI vs GT	159	0.743 (0.7, 0.788)	-0.001 (-0.027, 0.026)	<0.0001
Glass vs GT	159	0.745 (0.703, 0.786)		

Agreement for NASH diagnosis between reads on WSI and glass GT compared to reads on glass and glass GT were similar for all 3 pathologists (Table 28). For pathologist A, the difference between WSI reads and glass GT vs glass reads and glass GT was -0.006 (95% CI, -0.031, 0.0196). For pathologist B the difference between WSI reads and glass GT vs glass reads and glass GT was 0.0278 (95% CI, -0.034, 0.089) and the difference for pathologist C was -0.025 (95% CI, -0.069, 0.016).

Table 28: Agreement between reads on WSI and glass GT vs reads on glass and glass GT by Individual Pathologist

Pathologist	Comparison	N	Agreement Rate (95% CI)	Difference (95% CI)
A	WSI vs GT	159	0.843 (0.786, 0.899)	-0.006 (-0.031, 0.019)
A	Glass vs GT	159	0.849 (0.792, 0.906)	
B	WSI vs GT	158	0.633 (0.56, 0.707)	0.0278 (-0.034, 0.089)
B	Glass vs GT	157	0.605 (0.529, 0.679)	
C	WSI vs GT	159	0.755 (0.686, 0.824)	-0.025 (-0.069, 0.016)
C	Glass vs GT	159	0.780 (0.711, 0.843)	

Secondary Analysis

WKs between WSI read and glass read for each NASH component (Table 29) and each NASH component per pathologist were also determined (Table 30). For each NASH component, the overall WKs were higher than published values (Table 25). For NASH components per individual pathologist, pathologist A and C had the highest WKs across all score components and these WKs were all higher than published intra-pathologist Kappas (Table 25). For pathologist B, the WKs were lower than pathologist A and C but still in the published ranges listed in Table 30. For overall NAS score, all 3 pathologists WKs were higher than the published values from Davison 2020 (14) (Table 25), with WKs being the highest for pathologist A and C.

Table 29: Average WK between WSI reads and glass reads per NASH component

Feature	N	WK (95%CI)
Steatosis	159	0.882 (0.844, 0.916)
Lobular inflammation	159	0.761 (0.707, 0.809)
Hepatocellular ballooning	159	0.788 (0.732, 0.835)
Fibrosis	159	0.872 (0.837, 0.901)
NAS	159	0.795 (0.76, 0.825)

Table 30: WK between WSI reads and glass reads per NASH component by pathologist

Pathologist	Feature	N	WK **(95% CI)**
A	Steatosis	159	0.987 (0.966, 1)
	Lobular inflammation	159	0.938 (0.888, 0.981)
	Hepatocellular ballooning	159	0.926 (0.862, 0.972)
	Fibrosis	159	0.941 (0.903, 0.972)
	NAS	159	0.956 (0.93, 0.977)
B	Steatosis	157	0.741 (0.635, 0.83)
	Lobular inflammation	157	0.545 (0.441, 0.645)
	Hepatocellular ballooning	157	0.618 (0.498, 0.719)
	Fibrosis	153	0.733 (0.654, 0.803)
	NAS	157	0.602 (0.524, 0.668)
C	Steatosis	159	0.918 (0.87, 0.962)
	Lobular inflammation	159	0.801 (0.716, 0.877)
	Hepatocellular ballooning	159	0.821 (0.746, 0.885)
	Fibrosis	158	0.942 (0.905, 0.972)
	NAS	159	0.828 (0.779, 0.874)

4.3.11 Limitations

This AISight Clinical Trials platform validation study was performed on Aperio AT2 scanners at 40x magnification. Additional research may confirm generalizability of these findings for WSI from additional scanners and/or at different magnifications (e.g., 20x).

4.3.12 Discussion and Conclusions

This digital platform validation study demonstrates that the accuracy of NASH digital reads on the AISight Clinical Trials platform is non-inferior to reads performed with traditional light microscopy with glass slides. This is in line with studies performed for primary diagnoses by Leica (33) and Philips (34). This study demonstrated a significant non-inferior overall agreement of NASH assessment between WSI and glass GT reads vs glass and glass GT reads (NI margin of 0.05, difference of -0.001, 95% CI of (-0.027,0.026), and $p<0.0001$). Additionally, the agreement between WSI and glass GT reads vs glass and glass GT reads were shown to be similar for each individual participating pathologist. Average intra-reader WKs for each score component in this study were higher than WKs in published literature.

Varying level of intra-reader agreement was observed per pathologist per NASH component, which is expected as wide range of intra-reader WKs have previously also been shown in the literature (8,14,25). However, all 3 pathologists were still within the published ranges for intra-reader WKs, with 2 out of the 3 pathologists exceeding the published Kappas for all 4 NASH components (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis) and all 3 pathologists exceeding the WK for overall NAS score. It should be noted that despite an attempt to enroll 5-10% challenging cases (defined as NAS ≥ 4 with a score of 0 for at least one of the components (steatosis, lobular inflammation and ballooning), NAS = 4 with a score of > 1 for each of the components (steatosis, lobular inflammation and ballooning) or NAS = 3), actual enrollment based on ground truth scores resulted in approximately 40% challenging cases which may have contributed to variability in intra-reader agreements.

The results from this NASH digital platform validation study support the conclusion that the AISight Clinical Trials platform is non-inferior to the glass read in reference to glass GT, when used by pathologists in NASH trial

population diagnoses (defined as NAS $\geq$ 4 with a score of $\geq$ 1 for each component and absence of atypical features suggestive of non-NASH liver disease) and therefore can be utilized for NASH reads in clinical trials. Incorporating digital pathology into clinical trial workflows makes trial management more efficient, allows for multiple reads in parallel, and provides opportunities to utilize the most experienced pathologists on reader panels as geographic location is no longer a factor for selecting pathologists or shipping glass slides. Utilization of the AISight Clinical Trials platform will allow pathologists from all over the world to work on the same cases simultaneously and provide their results within hours of slide upload, shortening trial timelines, while allowing for accurate, gold standard assessments.

4.4 Validation of the AISight Translational Platform

4.4.1 Product Description

The AISight Translational platform is a RUO cloud-based software that enables PathAI to utilize WSIs for gathering annotations and scores for algorithm training and development, and in rare cases, PathAI partners for manual digital reads in retrospective clinical trials. The platform is a secure, web-based platform that has data entry capabilities, but no data analysis or long-term storage occurs. The data is extracted from the platform after the study completion and locked in eQMS. Digital pathology evaluations performed on the AISight Translational platform for clinical trials should not be used to inform patient care outside of the clinical trial and are for research purposes only.

The AISight Translational platform is a secure, web-based platform where each user can execute role-specific actions to upload, process, and view samples as slide images in DZI format. The platform allows for maximum flexibility in designing questions and data fields associated with each slide, making data collection faster and easier for algorithm development processes. The flexible configurability also allows the user to design specific case report forms (CRFs) for clinical studies.

The AISight Translational platform was utilized in collection of digital reads for ground truth in analytical and clinical validation, in collection of reads in overlay validation, and reference reads in analytical validation.

4.4.2 Objectives

The primary objective of the AISight Translational platform validation study was to validate the platform for NASH reads using glass slides scanned on the Aperio AT2 whole slide scanner by evaluating non-inferior agreement of NASH (defined as NAS $\geq$ 4 with a score of $\geq$ 1 for each component and absence of atypical features suggestive of non-NASH liver disease, similar to the definition used during NASH clinical trial enrollment) and non-NASH diagnosis between glass GT read and WSI read versus agreement between glass GT read and individual study pathologist glass read.

4.4.3 Study Design and Plan

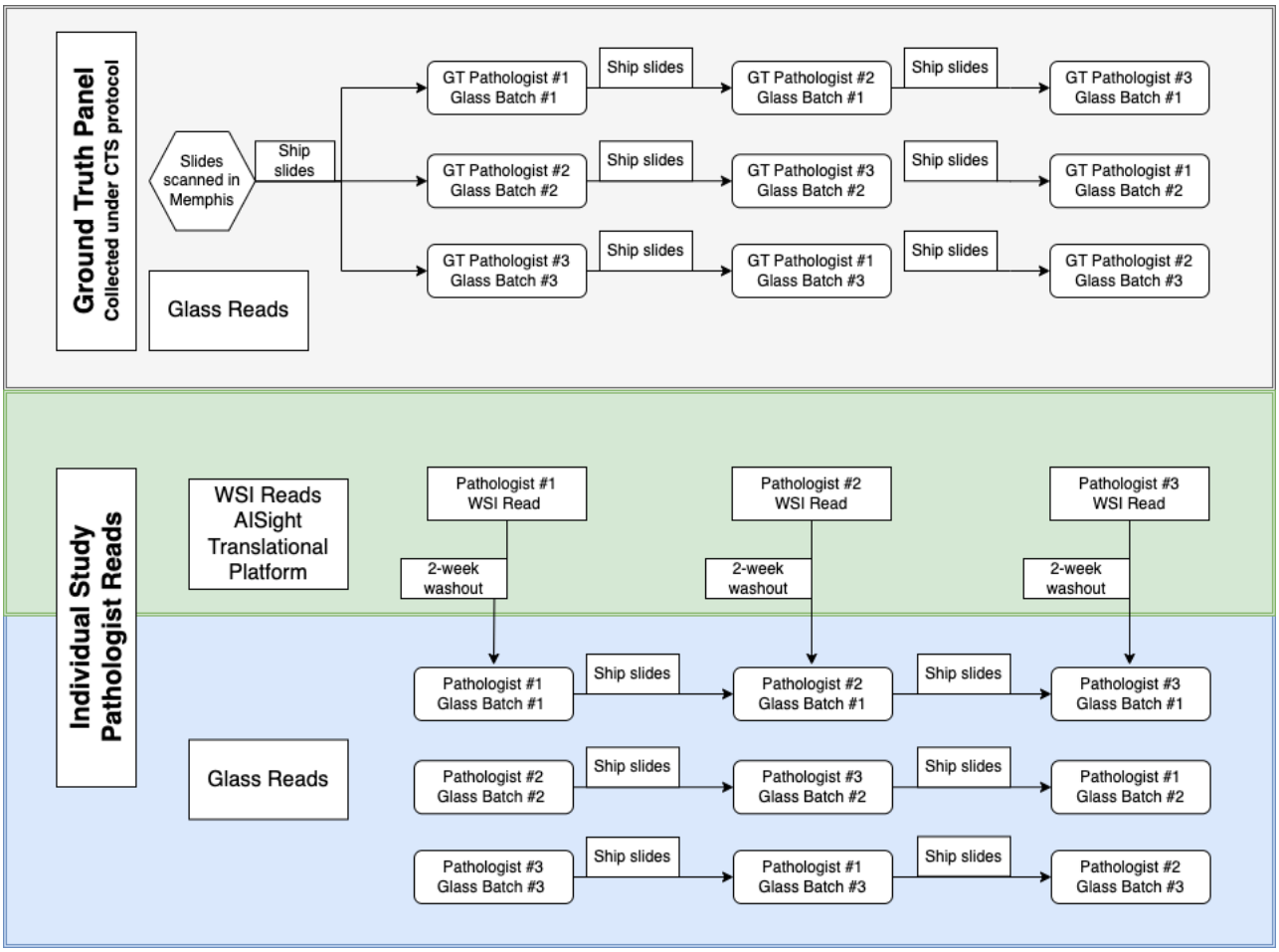
PathAI utilized already existing glass slides from completed Phase 2 clinical trials acquired from partners (screen failures from Phase 2B study from NorthSea Therapeutics NCT04052516 and enrolled cases from a Phase 2 study from Madrigal Pharmaceuticals NCT02912260) and from a third-party vendor (Precision for Medicine). All cases for this study include a H&E and a trichrome slide. The slides were previously scanned on Aperio AT2 scanner as part of the AISight Clinical Trials platform validation study described above.

AISight Translational platform reads were performed by 6 board-certified hepatopathologists who have demonstrated proficiency in reading manual NASH cases (see Appendix IIb for more details on documentation in PathAI's electronic Quality Management System (eQMS). The GT reads were collected on 160 glass slides

using a microscope by 3 board certified hepatopathologists from PathAI Contributor Network (collected under AISight Clinical Trials platform validation protocol described above in Section 4.1.4.2). These 3 pathologists had experience in reading NASH cases in their clinical practice and for PathAI projects and were the same as the GT pathologists for AISight Clinical Trials platform validation described above. The GT score was computed as the median of all 3 scores. Additionally, the majority of the 3 GT pathologists' responses were used to assess for presence of atypical features.

A different set of 3 board certified hepatopathologists performed the independent study reads. They read all cases twice, once on glass using a microscope and once on Aperio AT2 scanned WSIs using the AISight Translational platform. A minimum of 2-week washout was required between each set of reads. See Figure 18 for study read logistics.

Figure 18: Study logistics



The slide set consisted of 160 cases (320 slides - each case consists of one H&E slide and one trichome slide) from liver core needle biopsies. 60-70% of the cases came from NASH and 30-40% of the cases were from NAFLD and other (non-NAFLD) liver disorders, including but not limited to hepatitis B, hepatitis C, active hepatitis, viral hepatitis, cirrhosis, and intrahepatic cholestasis. Five to ten percent of the cases were chosen to be challenging, defined as NAS ≥4 with a score of 0 for at least one of the components (steatosis, lobular inflammation, and ballooning), NAS =4 with a score of ≥1 for each of the components (steatosis, lobular

inflammation, and ballooning) or NAS =3. For glass reads, the 160 cases were split into 3 batches and the pathologists read 1 batch at a time.

4.4.4 Dataset

Three hundred and twelve (312) slides (156 cases) were enrolled in the study by the GT pathologists. The glass slides were from completed Phase 2 studies (enrolled samples from Madrigal's non-cirrhotic phase 2 study for thyroid hormone receptor-b agonist and screen failures from NorthSea Therapeutics non-cirrhotic phase 2b study for carboxylic acid) and liver biopsy glass slides purchased from Precision for Medicine. Overall, 96 slides were enrolled from NorthSea, 184 slides were enrolled from Madrigal and 32 slides were enrolled from Precision for Medicine.

Table 31: AISight Translational Platform Validation Population

Source/Trial Phase	Drug Class	Enrollment Criteria
Madrigal MGL-3196 Phase 2 Enrolled population	Thyroid hormone receptor-β agonist	● NAS ≥ 4, with a score of at least 1 in each component (steatosis, lobular inflammation, and ballooning) ● Fibrosis stage 1 to 3
Northsea Icona Trial, Phase 2B Screen-failed population only	Carboxylic acid	● NAS ≥ 4, with a score of at least 1 in each component (steatosis, lobular inflammation, and ballooning) ● fibrosis score F1 to F3, inclusive (F1 capped at 30%)
Precision for Medicine	N/A	-

4.4.5 Selection of Study Population/ Cases

Inclusion Criteria

- Study slides scanned at pre-specified magnification (40x).
- Images are quality checked according to scanner instructions and can be rescanned if needed. All information logged by the lab technician and captured per lab established SOPs.
- Two (2) slides included per case: one H&E slide and one trichrome slide.
- One case per patient, i.e., unique cases.
- De-identified cases.
- Only liver biopsies included.

Exclusion Criteria

- Cases for which the slides do not fulfill quality check according to the scanning acquisition device.
- Cases with indelible markings.
- Cases with any patient identifying information.
- Cases with any tissue other than liver FFPE tissue.
- Cases without both H&E and trichrome slides available.

4.4.6 General Procedures

Blinding: All participating pathologists had their own unique log in to the AISight Translational platform and OpenClinica electronic data capture (eDC) platform and were assigned their specific study cases. The pathologists were blinded to each other's assessments and to their own assessments from different modalities (glass and digital

reads). All PathAI staff (except for the unblinded clinical data managers and unblinded clinical scientist) involved in this study were blinded to the data until the database was locked. No PHI data was collected in this study.

Glass Slides Scanning and Handling: Glass slides utilized for this validation study were from completed Phase 2 studies (enrolled samples from Madrigal Pharmaceuticals MGL-3196 study NCT02912260, screen failures from NorthSea Therapeutics ICONA clinical study NCT04052516) and glass slides purchased from Precision for Medicine, a commercial biobank that collects their samples under IRB oversight following the highest industry standards. These slides were labeled and scanned by PathAI Biopharma lab on Aperio AT2 whole slide scanner at 40X magnification (under IRB protocol number PATHAI-090822-016). The Biopharma Lab was also responsible for shipping the glass slides to the 6 participating pathologists and generating shipping labels for each shipment. The lab also re-labelled the glass slides with the original trial information after the completion of the study.

4.4.7 Pathologist Training

All pathologists were trained on study protocol and required tasks (See Appendix VIc for Study Case Report Forms) by the principal investigator (PI) prior to participating in any study activities. All pathologists also signed an attestation form acknowledging the completion of training. All training records are stored in PathAIs eQMS.

4.4.8 Data Handling

All data was entered electronically in the AISight Translational platform for the digital reads and OpenClinica for the glass reads. After the completion of the study, all data was securely downloaded from the platforms and stored in the clinical data management S3 bucket. Data for analysis was uploaded to PathAIs eQMS) once the database lock form was approved. PathAI designated clinical data managers and a clinical scientist who were unblinded to the data and had access to all study information for the purpose of monitoring data and resolving any queries. Data monitoring and querying was performed as described in the Data Management Plan.

Data was downloaded from the database service for the AISight Translational platform and OpenClinica over the course of the study for data monitoring purposes. All relevant study data along with corresponding documentation was uploaded to the Clinical Data Management Amazon Web Services (AWS) bucket which only unblinded clinical data managers have access to.

4.4.9 Statistical Methods and Determination of Sample Size

Primary: The primary endpoint is the agreement of NASH (defined as NAS $\geq$ 4 with a score of $\geq$ 1 for each component and absence of atypical features suggestive of non-NASH liver disease, similar to the definition used during NASH clinical trial enrollment) and non-NASH diagnosis between glass GT read and WSI compared to the agreement of NASH and non-NASH diagnosis between glass GT read and individual study pathologist glass read.

The null hypothesis is that the agreement of NASH and non-NASH diagnosis between glass GT read and WSI is inferior to the agreement of NASH and non-NASH diagnosis between glass GT read and individual study pathologist glass read discounted by a non-inferiority margin of 0.05. The alternative hypothesis is that the agreement between glass GT read and WSI is non-inferior to the agreement between glass GT read and individual study pathologist glass reads discounted by a non-inferiority margin of 0.05. These hypotheses are stated as follows:

$$H0: \pi_{GD} \leq \pi_{GG} —0.05$$

$$Ha: \pi_{GD} > \pi_{GG} —0.05$$

where π_{GD} is the average agreement across 3 pathologists of NASH and non-NASH diagnosis between glass GT read and WSI and π_{GG} is the average agreement across 3 pathologists of NASH and non-NASH diagnosis between glass GT read and individual study pathologist glass reads.

π_{GD} will be shown to be statistically non-inferior to 0.05 less than π_{GG} if Bootstrap percentile $p < 0.025$. Analysis was done on observed data and separately for each scanner.

Secondary: This endpoint consists of comparing study pathologist scores for the four primary NASH components (NAS components on H&E and CRN fibrosis on trichrome slides), and the overall NAS score on WSI and glass reads. This endpoint was evaluated as described below:

Linearly weighted Kappa concordance statistics between glass and WSI read for each of the pathologists, each of the NASH components, and overall NAS score will be computed. Overall, linearly weighted Kappa for each NASH component and overall NAS score will be computed by averaging the WK for the three (3) pathologists. Bootstrap 95% confidence intervals will be provided on the overall as well as per pathologist linearly weighted Kappa. These concordance estimates will be compared to the published range in Table 25 The secondary endpoints will be analyzed separately for each scanner.

These analyses are based on observed data.

Determination of Sample Size
The College of American Pathologists (CAP) guidelines recommend a minimum of 60 cases to ensure that diagnostic performance based on digitized slides is at least equivalent to that of glass slides and light microscopy and to identify and rectify risks associated with the technology (29). With substantial inter-rater variability in NASH scoring and diagnosis, a non-inferiority design was determined to be more appropriate for the NASH trial population than a direct comparison of agreement between glass and digital reads. A sample size of 160 slides was selected to provide a degree of precision around the estimates and to account for not evaluable slides, and any incidental breakage of glass slides.

4.4.10 AISight Translational Platform Validation Results

One hundred and fifty-six (156) cases were enrolled in the study by the GT pathologists. The glass slides were from completed Phase 2 studies (enrolled samples from Madrigal Pharmaceuticals MGL-3196 study NCT02912260 and screen failures from NorthSea ICONA study NCT04052516) and liver biopsy glass slides purchased from Precision for Medicine.

No demographic information for the slides enrolled in the study is available. The represents both failed and enrolled NASH clinical trial patient populations, liver biopsies from other liver diseases (including but not limited to hepatitis B, hepatitis C, active hepatitis, viral hepatitis, cirrhosis, and intrahepatic cholestasis) and normal liver. The dataset also contains variability in sample staining (including performed by multiple collection/preparation sites). Distribution of slides based on slide level score from glass GT are listed in
Table 32.

We aimed to include 5-10% of challenging cases, defined as NAS ≥ 4 with a score of 0 for at least one of the components (steatosis, lobular inflammation, and ballooning), NAS = 4 with a score of > 1 for each of the

components (steatosis, lobular inflammation and ballooning) or NAS = 3. Cases were selected with previously captured individual pathologist scores. However, considering inter-pathologist agreement for NAS can be low to moderate (14) (
Table 32.), especially for borderline cases, it was not surprising that, based on the study glass GT, the slide set included 39.1% of cases that met the definition for challenging cases.

Table 32: Distribution of Slides Based on Glass GT

Feature	Score	% (n/N)
Steatosis	0	8.3 (13/156)
	1	32.7 (51/156)
	2	30.1 (47/156)
	3	28.8 (45/156)
Lobular inflammation	0	1.9 (3/156)
	1	61.5 (96/156)
	2	34.6 (54/156)
	3	1.9 (3/156)
Hepatocellular ballooning	0	23.1 (36/156)
	1	56.4 (88/156)
	2	20.5 (32/156)
Fibrosis	0	7.1 (11/156)
	1	28.2 (44/156)
	2	29.5 (46/156)
	3	27.6 (43/156)
	4	7.7 (12/156)
NAS	0	0.6 (1/156)
	1	6.4 (10/156)
	2	10.3 (16/156)
	3	19.2 (30/156)
	4	17.3 (27/156)
	4; at least 1 score of 0	2.6 (4/156)
	5	21.8 (34/156)
	6	12.8 (20/156)
	7	8.3 (13/156)
	8	0.6 (1/156)

Primary Analysis

One hundred and fifty-six (156) cases were enrolled in the study by 3 GT pathologists by reading glass slides using a microscope. A separate set of 3 study pathologists evaluated the same 156 cases once as glass slides using a microscope, then as WSIs on AISight Translational platform on slides scanned on Aperio AT2 scanner, with a minimum of 2-week washout between different modalities. Each pathologist read the entire slide set on WSIs first and after 2-week washout they read the entire slide set on glass.

The acceptance criteria for non-inferiority (with a margin of 0.05) agreement for NASH diagnosis between reads on WSI and glass GT compared to reads on glass and glass GT for slides scanned on Aperio AT2 scanner was met with a difference of -0.004 (95% CI of (-0.045, 0.036); p=0.0110;

Table 33). The agreement between study pathologists reads on Slides using WSIs scanned on Aperio AT2 scanner and glass GT was 0.788 (95% CI, 0.739, 0.838) and the agreement for glass reads and glass GT was 0.793 (95% CI, 0.748, 0.838).

Table 33: Agreement between reads on WSI and glass GT vs reads on glass and glass GT for slides scanned on Aperio AT2

Modality	N	Agreement Rate (95% CI)	Difference (95% CI)	P value
Glass vs GT	156	0.793 (0.748, 0.838)	-0.004 (-0.045, 0.036)	0.0110
WSI vs GT	156	0.788 (0.739, 0.838)		

For each individual pathologist, for slides scanned on the Aperio AT2 scanner, the agreement for NASH diagnosis between reads on WSI and glass GT compared to reads on glass and glass GT were similar for all 3 pathologists (Table 34). For pathologist A, the difference between WSI reads and glass GT vs glass reads and glass GT was -0.026 (95% CI, -0.09, 0.045). For pathologist B the difference between WSI reads and glass GT vs glass reads and glass GT was 0.0513 (95% CI, -0.016, 0.122) and the difference for pathologist C was -0.038 (95% CI, -0.103, 0.026).

Table 34: Agreement between reads on WSI and glass GT vs reads on glass and glass GT by individual pathologist for slides scanned on Aperio AT2 scanner

Pathologist	Modality	N	Agreement Rate (95% CI)	Difference (95% CI)
A	Glass vs GT	156	0.769 (0.705, 0.833)	-0.026 (-0.09, 0.045)
	WSI vs GT	156	0.744 (0.673, 0.814)	
B	Glass vs GT	156	0.750 (0.679, 0.814)	0.051(-0.016, 0.122)
	WSI vs GT	156	0.801 (0.737, 0.865)	
C	Glass vs GT	156	0.859 (0.808, 0.91)	-0.038 (-0.103, 0.026)
	WSI vs GT	156	0.821 (0.756, 0.878)	

Secondary Analysis

WKs between WSI read and glass read for each NASH component (

Table 35) and each NASH component per pathologist were also determined (Table 36). For slides scanned on Aperio AT2 scanner, for each NASH component, the average WK was in the range of published values (Table 25). All 3 pathologists' individual WKs for each NASH component were either in the range of published literature or close to the lower bound of the published intra-reader WKs (Table 25). For overall NAS score, all 3 pathologists were higher than the published WK value of 0.372 from Davison 2020 (14) (Table 25), with WKs being similar for all 3 pathologists.

Table 35: Average WK between WSI reads and glass reads per NASH component for slides scanned on Aperio AT2 scanner

Feature	N	WK (95% CI)
Steatosis	156	0.811 (0.761, 0.854)
Lobular inflammation	156	0.440 (0.339, 0.519)
Hepatocellular ballooning	156	0.591 (0.51, 0.661)
Fibrosis	156	0.711 (0.655, 0.760)
NAS	156	0.652 (0.601, 0.695)

Table 36: WK between WSI reads and glass reads per NASH component by pathologist for slides scanned on Aperio AT2 scanner

Pathologist	Feature	N	WK (95% CI)
A	Steatosis	156	0.811 (0.746, 0.869)
	Lobular inflammation	156	0.515 (0.359, 0.652)
	Hepatocellular ballooning	156	0.634 (0.504, 0.742)
	Fibrosis	156	0.632 (0.539, 0.714)
	NAS	156	0.675 (0.603, 0.739)
B	Steatosis	156	0.802 (0.729, 0.871)
	Lobular inflammation	156	0.351 (0.227, 0.465)
	Hepatocellular ballooning	156	0.579 (0.476, 0.678)
	Fibrosis	155	0.704 (0.63, 0.768)
	NAS	156	0.622 (0.551, 0.685)
C	Steatosis	156	0.819 (0.750, 0.879)
	Lobular inflammation	156	0.454 (0.286, 0.592)
	Hepatocellular ballooning	156	0.560 (0.440, 0.674)
	Fibrosis	144	0.798 (0.724, 0.86)
	NAS	156	0.657 (0.592, 0.719)

4.4.11 Limitations

This AISight platform validation study was performed on Aperio AT2 scanners at 40x magnification. Additional research may confirm generalizability of these findings for WSI from additional scanners and/or at different magnifications (eg. 20x).

4.4.12 Discussion and Conclusions

This digital platform validation study demonstrates that NASH digital reads on the AISight Translational platform for slides scanned on Aperio AT2 scanner are non-inferior to reads performed with traditional light microscopy with glass slides. This is in line with studies performed for primary diagnoses by Leica (33) and Philips (34). This study demonstrated a significant non-inferior overall agreement of NASH

assessment between WSI and glass GT reads vs glass and glass GT reads (NI margin of 0.05, difference of -0.004, 95% CI of (-0.045,0.036), and p=0.0110). In addition, the agreement of digital reads with GT and of glass reads with GT were shown to be similar for each individual participating pathologist (pathologist A 0.769 and 0.744, pathologist B 0.750 and 0.801 and pathologist C 0.859 and 0.821). Overall and per pathologist intra-reader, inter-modality (glass to digital) WKs for each score component in this study were within the range or close to the lower bound of published WKs, with NAS intra-rater agreed being higher for all pathologists in this study compared to the published values from Table 25 (published WK range for steatosis 0.666-0.83, for lobular inflammation 0.227 – 0.60, for hepatocellular ballooning 0.32-0.66, for fibrosis 0.64-0.85, for NAS value 0.372; (8,14,25). The overall WKs compared to GT were slightly different between platforms, but these values were all within the expected inter-reader range for each histologic component and for total NAS. The variation can likely be attributed to the utilization of different panel readers for each validation study and the known variability between readers, as demonstrated in the literature (8,14,25). Importantly, glass to GT and digital to GT read agreements were equivalent for both platforms. Varying levels of intra-reader agreement was also observed per study pathologist glass and Aperio AT2 WSI read per NASH component, which is expected as a wide range of intra-reader WKs have previously also been shown in the literature (8,14,25). For Aperio AT2 scanned WSIs, pathologists B and C were within the published ranges for intra-rater Kappas, and pathologist A was slightly below for one of the score categories (Kappa for fibrosis for pathologist A was 0.632 vs 0.64 the lowest range in literature).

It should be noted that despite an attempt to enroll 5-10% challenging cases based on single pathologist scores, actual enrollment based on ground truth scores resulted in approximately 40% of cases meeting the definition of borderline or challenging which may have contributed to higher variability in intra-reader agreements. Despite this over-enrichment for challenging cases the study population, the high variability seen in this study is consistent with published literature describing inter-pathologist agreement rates for NAS score of approximately 30% (22).The results from this NASH digital platform validation study supports the conclusion that the AISight Translational platform is non-inferior to the glass read in reference to glass GT for slides scanned on Aperio AT2 scanner when used by pathologists in NASH trial population diagnoses (defined as NAS > 4 with a score of > 1 for each component and absence of atypical features suggestive of non-NASH liver disease) and therefore can be utilized for NASH reads in clinical trials and validation studies.

Incorporating digital pathology into clinical trial workflows makes trial management more efficient, allows for multiple reads in parallel, and provides opportunities to utilize the most experienced pathologists on reader panels as geographic location is no longer a factor for selecting pathologists. Additionally, with the ongoing development and validation of digital pathology tools including machine learning algorithms, these digital platforms have the potential to enhance a pathologist's evaluation of histology in drug development and in the clinic. Utilization of the AISight Translational platform will allow pathologists from all over the world to work on the same cases simultaneously and provide their results within hours of slide upload, shortening trial timelines, while allowing for accurate, gold standard assessment.

4.5 Analytical Validation

4.5.1 Study Purpose

The purpose of this study is to generate evidence of the precision and accuracy of AIM-NASH in measuring each component of the NAS score (steatosis, lobular inflammation, and hepatocellular ballooning) and CRN fibrosis stage.

4.5.2 Objectives

Primary Objectives

- Accuracy: To evaluate for non-inferior agreement, we calculated linearly weighted Kappa for AIM-NASH scores vs. GT and compared that to mean pairwise Kappa for IMR vs. GT for all NASH components.
- Reproducibility: To evaluate for superior performance to published manual pathologist scoring in terms of Agreement Rate between AIM-NASH scoring on whole slide images (WSIs) scanned from the same slide by 3 different operators and 3 different Aperio AT2 (Leica) scanners. Manual inter-pathologist agreement per histologic component is determined to be less than 85% based on published literature, therefore we aim to show statistical superiority above this threshold.
- Repeatability: To evaluate for superior performance to published manual pathologist scoring in terms of Agreement Rate between AIM-NASH scoring on WSIs scanned from the same slide by a single operator and single Aperio AT2 scanner at three separate inter-day times. Manual intra-pathologist agreement per histologic component is determined to be less than 85% based on published literature, therefore we aim to show statistical superiority above this threshold.

Secondary Objectives

- To provide accuracy within known clinical subsets (e.g., different visit: screening, baseline, post-baseline) and aggregate NASH component scores for the following:
    - CRN fibrosis F4 vs other
    - CRN fibrosis F0&F1 vs other
    - NAS aggregate score $\geq$4 vs other (NAS aggregate score defined as the sum of scores for steatosis, lobular inflammation, and hepatocellular ballooning).
- Secondary endpoints for reproducibility and repeatability include to provide analyses within known treatment time points (screening, baseline, and post-baseline).

Exploratory Objectives

- To provide summary statistics reporting distributions of trial of origin, timepoint, baseline co-morbidities (Type-2 Diabetes) and NASH treatment for all slides utilized in AV.
- To provide accuracy, reproducibility, and repeatability analyses in clinical subsets defined by each trial of origin and per score component. For repeatability and reproducibility, per-score agreement across operators, scanners, and dates.

4.5.3 Study Design and Plan

The AV for AIM-NASH is a multi-site instrument precision and accuracy study. AV tests the performance of AIM-NASH itself, without pathologist review of scores. The full AIM-NASH workflow (AI-assisted) with pathologist review is performed in clinical validation (CV).

Analytical validation of AIM-NASH was performed against glass slides provided by sponsors and selected from completed phase 2 (Bristol Myers Squibb FALCON 1 trial NCT03486899 and FALCON 2 trial NCT03486912) and phase 3 NASH trials (Intercept Pharmaceuticals REGENERATE trial). The data sources used for AV contain a broad spectrum of disease presentation, represent both screened and enrolled patient populations, including study subjects who may have regressed or progressed during a clinical trial from both placebo and treatment groups, and reflect the NASH clinical trial population. The dataset also contained variability in sample collection (historical biopsies vs. study biopsies), staining (including performed by multiple collection/preparation sites), and populations representing various treatments with candidate therapies. We aimed to choose each NAS component and fibrosis stage based on Table 37 and for each score component to be represented at least between 10-40% in the dataset (based on existing AIM-NASH scores). The purpose of including this broad spectrum of samples across collection and handling conditions, as well as disease presentation and clinical trial settings, was to allow for representation of data and populations that are reflective of the variability encountered in NASH trials where this biomarker would be used. Liver biopsy slides distinct from those used for training and verification of AIM-NASH were used in AV.

Table 37: Planned Score Distribution for AV

Feature	Score	Percentage of Cases	
		Accuracy, N=600 % (n)	**Repeatability and Reproducibility, N=150 % (n)**
Steatosis	0	20 (120)	20 (30)
	1	20 (120)	20 (30)
	2	30 (180)	30 (45)
	3	30 (180)	30 (45)
Lobular inflammation	0	20 (120)	20 (30)
	1	20 (120)	20 (30)
	2	30 (180)	30 (45)
	3	30 (180)	30 (45)
Hepatocellular ballooning	0	20 (120)	20 (30)
	1	40 (240)	40 (60)
	2	40 (240)	40 (60)
Fibrosis	0	15 (90)	15 (22.5)
	1	15 (90)	15 (22.5)
	2	25 (150)	25 (37.5)
	3	25 (150)	25 (37.5)
	4	20 (120)	20 (30)

Accuracy

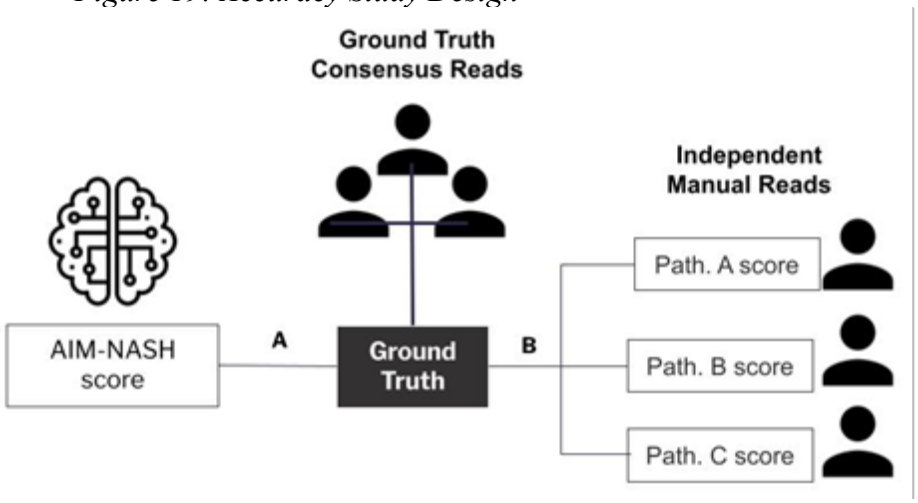
Accuracy of AIM-NASH is assessed relative to that of a practicing board certified liver pathologist (Figure 19). Accuracy was assessed separately for each of the 4 NASH components (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis). A single AIM-NASH score, a GT and IMRs from a minimum of 3 pathologists were collected for each case. Six hundred cases (1200 total slides; each case consists of a H&E slide and a trichrome slide) were planned to be enrolled in the study. The GT and IMRs were performed on AISight Translational platform and AIM-NASH was run on AISight Clinical Trials platform.

The GT was established by two panels of 2 expert liver pathologists with a third pathologist serving as tiebreaker. The pathologists were chosen based on their previous experience and results of previously completed slides for PathAI. All slides from AV and CV were split between the two panels, so that each panel read about half of the overall dataset (AV and CV slides combined).

In cases where the two primary readers disagreed with the score on any of the NASH components (steatosis, lobular inflammation, hepatocellular ballooning and/or fibrosis), the slide was sent out to the third tiebreaker pathologist. To blind the tiebreaker to the discordant component, all NASH components were sent out for scoring. The tiebreaker pathologist was blinded to the scores of the two primary pathologists. If the tiebreaker pathologist agreed with one of the primary pathologists for the originally discordant component, this was then the final score for that NASH component. If the third pathologist disagreed with both primary pathologists, a joint panel call was held with the three pathologists to come to a consensus, with the tiebreaker providing the final score in the rare case that consensus was not reached. The tiebreaker pathologist was the same for both panels. Overall, 5 pathologists provided scores for GT. These 5 pathologists were unique and not used for IMRs.
IMRs were performed by 8 qualified PathAI Contributor Network liver pathologists. Each AV slide was read by a minimum of 3 pathologists. Pathologists were selected by their previous experience in NASH trials and/or clinical experience with NASH, as well as performance tested to ensure proficiency. IMRs who scored the same slide did not have to come to a consensus, so no panel calls were held. For IMRs, not all slide pairs (H&E and trichrome) for each case were read by the same pathologist.

All slides for accuracy study were scanned at the Covance Indianapolis site on Aperio AT2 scanner at 40x magnification following established lab Standard Operating Procedures (SOPs). After scanning was complete, Covance lab technicians uploaded the WSIs onto AISight Clinical Trials platform and after AIM-NASH run, finalized the cases on the AISight Clinical Trials platform.

Figure 19: Accuracy Study Design



Repeatability

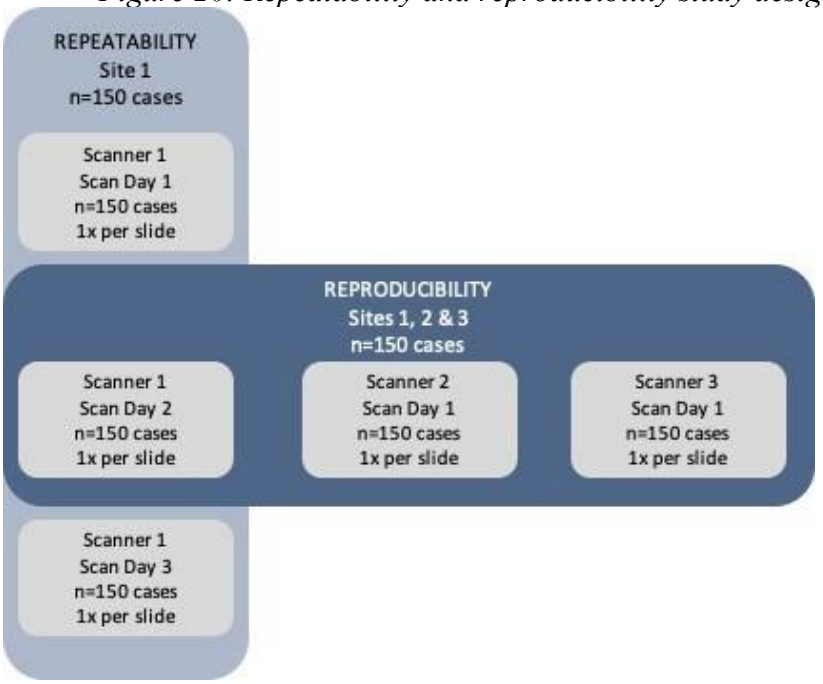
For evaluation of repeatability, a subset (n=150) of accuracy slides were scanned multiple times using the same Aperio AT2 scanner at 40x magnification on 3 non-consecutive days at Covance Perimeter Park lab (Figure 20). The scanner used was qualified and images were QC'd according to Covance SOPs before use. The 150 cases were chosen to have a similar score distribution to the accuracy dataset. Similar to accuracy, after scanning was complete, Covance lab technicians uploaded the WSIs onto AISight Clinical Trials platform and after AIM-NASH run, finalized the cases on the AISight Clinical Trials platform.

Reproducibility

For assessment of reproducibility, the same subset (n=150) of accuracy slides that was used in repeatability, were scanned once at 3 different Covance labs by 3 different operators using 3 different Leica AT2 scanners at 40x magnification. The 3 scans utilized in reproducibility were defined as follows (Figure 20).

- The "repeatability site" contributed scans performed on Day 2 to the reproducibility.
- The "accuracy site" contributed scans performed for the accuracy study to the reproducibility.
- The "reproducibility site" contributed scans performed for the reproducibility study specifically to the reproducibility.

Figure 20: Repeatability and reproducibility study design



4.5.4 Dataset

Liver biopsy slides distinct from those used for training and validating the AIM-NASH model algorithm are obtained to support analytical validation. One (1) trichrome and one (1) H&E slide will comprise a subject level case. Glass slides were sourced to reflect different clinical trials (including staining variability), phases (P2, P3, etc.), visits (screening, pre-, post-treatment), labs and other real-world heterogeneity. Where available, slide, tissue and clinical meta-data were used to guide selection and balancing of slides for AV. This sample selection strategy is intended to provide a generalizable dataset comprised of multiple clinical trials, conditions, and sample types that are broadly representative of the variability encountered in NASH trials where this biomarker would be used. These glass slides were sourced from completed clinical trials and a chain of custody will be maintained while handling the samples (following Good Clinical Practice Guidelines). Where prior NASH assessments are not available, the AIM-NASH algorithm will be used to estimate scores for balancing purposes (See Analysis Sets and Subgroups).

A final slide set was chosen to ensure coverage of meta-data features, balance across expected NASH assessment scores and predefined sample size requirements for Accuracy, Repeatability and Reproducibility.

Table 38: Analytical Validation Population

Trial Name and Sponsor	Trial Phase	Drug/Drug Class	Enrollment Criteria	Total Available Sample Size from Original Trial
REGENERATE Intercept Pharmaceuticals	3	Obeticholic Acid/ Farnesoid X receptor agonist	Presence of all 3 NAS components Fibrosis stage 2 or stage 3 OR Fibrosis stage 1a or stage 1b if accompanied by ≥1 of the following risk factors: Obesity (BMI ≥30 kg/m2) Type 2 diabetes diagnosed per 2013 American Diabetes Association criteria ALT >1.5× upper limit of normal (ULN).	6000
FALCON 1	2	Pegbelfermin / Pegylated FGF21 (anti-fibrotic)	A score of at least 1 for each NAS component Fibrosis stage 3	250
FALCON 2 Bristol Myers Squibb	2	Pegbelfermin / Pegylated FGF21 (anti-fibrotic)	Biopsy must be consistent with NASH Biopsy must be consistent with cirrhosis (stage 4)	281

4.5.5 Selection of Study Population/ Cases

Inclusion Criteria

- Slide is listed on master AV manifest at Covance.
- Accession contains both an H&E and trichrome slide.
- De-identified NASH biopsy glass slides (H&E and trichrome-stained slides from phase 2 and phase 3 NASH clinical trials) from screening and post-treatment populations, representing a wide spectrum of disease (characterized by NAS score and CRN fibrosis stage), a variety of staining (performed in different labs), multiple time points of collection (baseline, follow-up), and from trials with different classes of drug targets.

Exclusion Criteria

- Slide is broken or damaged at the time of receipt at Covance site where scanning was performed.
- Case contains only H&E or only trichrome slide
- Slide has been deemed "not evaluable" by manual read
- Any trichrome stain except for trichrome blue

4.5.6 General Procedures

Blinding

All participating pathologists had their own unique log in to the AISight Translational platform and they were assigned jobs with relevant slides. All study pathologists were blinded to trial source, timepoint, or treatment arm,

as well as each other's individual scores. This includes the GT and IMRs. Pathologists were not presented with AIM-NASH algorithm scores. All PathAI staff (except for the unblinded clinical data managers and unblinded clinical scientists) involved in this study were blinded to the data until the database was locked.

Glass Slides Scanning and Handling

All glass slide handling (including accessioning, shipping, labeling if needed, and storage) and scanning was performed by 3 Covance sites. Each site followed their internal SOPs for slide shipping, scanning, and handling. In addition, all chain of custody and scanning information was entered into OpenClinica electronic Data Capture (eDC) platform by Covance lab technicians. All Covance staff handling slides for this study were trained on the study protocol prior to the start of the study. All training records for the Covance staff are stored in PathAI's eQMS.

4.5.7 Pathologist Training

All GT and IMRs pathologists received and were asked to review CRN NASH histology scoring guidelines and corresponding literature prior to completing any NASH scoring tasks. They received separate instructions for each H&E and trichrome scoring task. These instructions detailed how many slides were to be reviewed and the type of review needed. These instructions also included the CRN NASH scoring criteria.

4.5.8 Data Handling

All pathologists entered their data electronically in the AISight Translational platform and AIM-NASH was run on the AISight Clinical Trials platform. After the completion of the study, all data was securely downloaded from the platforms and stored in the clinical data management's Amazon Web Services (AWS) bucket. Data for analysis was uploaded to the PathAI's eQMS system after the database lock.

All data entered by the Covance labs was entered into the OpenClinica eDC. Covance scanning technicians accessing the AISight Clinical Trials platform with their own logins were able to finalize cases but did not perform any score review or annotations of exclusion regions. PathAI designated clinical data managers and clinical scientists who were unblinded to the data and had access to all study information for the purpose of monitoring data and resolving any queries.

Data Quality Assurance

PathAI utilizes SOPs designed to ensure that research procedures and documentation are consistently conducted/prepared to the highest quality standards. These SOPs also require compliance with ICH Guideline for Good Clinical Practice (E6 (4.6.4)). Additionally, the contracted Covance sites have established SOPs in place to maintain quality standards.

4.5.9 Statistical Methods and Determination of Sample Size

Primary Analysis

Accuracy: Accuracy of AIM-NASH is assessed separately for each of the 4 NASH component scores (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis). A single AIM-NASH score, a single GT score, and a set of at least 3 IMR scores are collected for each NASH case component. As a reference, mean (across

IMRs) pathologist-GT concordance will be quantified by the linearly WK statistic for each NASH component (represented by "B" in
Table 39; for expected values, see Table 40). AIM-NASH agreement with GT will be quantified by linearly WK between AIM-NASH and GT for each NASH component (represented by "A" in
Table 39). To show agreement non-inferior to a standard qualified pathologist, AIM-NASH concordance with GT will be shown to be significantly greater (Bootstrap percentile $p < 0.025$) than 0.1 less than the mean IMR concordance with GT for each NASH assessment (H1: $A > (B - 0.1)$). Overall acceptance will be determined by meeting this non-inferiority criteria for all 4 NASH assessments (steatosis, ballooning, lobular inflammation, and CRN fibrosis stage). If non-inferiority is shown, then accuracy will also be tested for superiority of AIM-NASH concordance with GT to IMR concordance with the GT.

Repeatability: Repeatability of AIM-NASH is assessed separately for each of the 4 NASH component scores (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis). Three (3) AIM-NASH scores are collected for each NASH case component from slides scanned by the same Aperio AT2 scanner on 3 non-consecutive days.
To show repeatability superior to published intra-pathologist agreement across repeat assessment, mean Agreement Rate will be shown to be significantly greater (Bootstrap percentile $p < 0.025$) than 0.85 (H1: $R > 0.85$), an agreement rate exceeding known intra- pathologist performance across all NASH scores. Percent agreement is defined as specific similarity between two sets of paired, ordinal assessment scores will be quantified by agreement rate, defined by the proportion of matching assessment scores in the paired set.

Reproducibility: Reproducibility of AIM-NASH is assessed separately for each of the 4 NASH component scores (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis). Three (3) AIM-NASH scores are collected for each NASH case component from slides scanned by 3 different operators and 3 different Aperio AT2 scanners.
To show reproducibility superior to published inter-pathologist agreement across repeat assessments, mean Agreement Rate will be shown to be significantly greater (Bootstrap percentile $p < 0.025$) than 0.85 (H1: $R > 0.85$), an agreement rate exceeding known intra- pathologist performance across all NASH scores. Percent agreement is defined as specific similarity between two sets of paired, ordinal assessment scores will be quantified by agreement rate, defined by the proportion of matching assessment scores in the paired set.

Secondary Analysis

Accuracy: Accuracy within known clinical subsets (e.g., different visit: screening, baseline, post-baseline) and aggregate NASH component scores for the following will be computed:

1. CRN fibrosis F4 vs other
2. CRN fibrosis F0&F1 vs other
3. NAS aggregate score ≥ 4 vs other (NAS aggregate score defined as the sum of scores for steatosis, lobular inflammation, and hepatocellular ballooning).

For accuracy, Kappa concordance values for these comparisons will be represented separately for AIM-NASH vs GT s and the average of each IMR vs GT. 95% confidence intervals will be obtained from the bootstrap percentile method.

Justification of Non-Inferiority Margin (relevant to both AV and CV)

A non-inferiority margin of 0.1 for Linearly Weighted Kappa was chosen due to multiple different factors, outlined below.

1. Qualitative interpretation of Kappa: Linearly weighted kappa is a measure of ordinal class concordance where a value of 0 represents chance and 1 represents perfect agreement. Kappa is typically interpreted qualitatively in ranges of 0.2, where values of 1-0.8 represent complete agreement, 0.8-0.6 represents strong agreement, 0.6-0.4 moderate agreement, 0.4-0.2 weak agreement, and 0.2-0 chance agreement.

Therefore, for our chosen non-inferiority margin of 0.1, kappa represents half-width of these qualitative bins.

2. Range of observed inter-pathologist Kappas. Pairwise inter-pathologist Kappas were measured for the held-out test set, across three qualified, expert pathologists (N=220 trichrome, N=231 H&E). Kappa ranges across these three pairs were measured [Range (Min, Max)] as 0.1 (0.61, 0.71), 0.08 (0.41, 0.49), 0.15 (0.3, 0.45), and 0.18 (0.48, 0.66) for fibrosis, lobular inflammation, ballooning, and steatosis, respectively.

The non-inferiority margin of 0.1 kappa conforms to these ranges of inter-pathologist concordance for fibrosis, ballooning, and steatosis. Although a non-inferiority endpoint is being employed, it is possible that the data may show some improvement (for particular score components) in agreement rates between the AIM-NASH tool and ground truth compared to inter-pathologist Kappas published by the NASH CRN. However, improvement compared to published values intra- and inter-rater variability will be demonstrated across all score components during reproducibility studies, where the endpoint requires greater than 85% agreement across scans and scanning operators/sites. If achieved, then this demonstrates superiority compared to relevant published Kappa values. The combination of non-inferior accuracy and superior reproducibility (see sections directly below) should ensure accurate, consistent scoring during enrolment and improved calculation of change over time for histologic-based endpoint analyses.

Repeatability and Reproducibility

Secondary endpoints for reproducibility and repeatability include analysis within known treatment time points (screening, baseline, and post-baseline).

For repeatability and reproducibility, agreement rates along with 95% confidence intervals obtained from the bootstrap percentile method will be presented for each NASH feature across baseline and post-baseline timepoints.

Exploratory Endpoints

Summary statistics reporting distributions of trial of origin, timepoint, baseline comorbidities (Type-2 Diabetes) and NASH treatment for all slides utilized in AV will be provided.

Accuracy, reproducibility, and repeatability analyses in clinical subsets defined by each trial of origin and per score component will be provided.

Table 39: Pathologist groups, pathologist Kappas and AI-assisted Kappas

NASH Assessment	NASH Component Score	Concordance Comparisons	Definitions
Ground truth (GT)	GT	-	-
Independent Manual Reads (IMR)*	IMR1, IMR2, IMR3, IMR4, IMR5, IMR6, IMR7, IMR8	IMR1 v GT, IMR2 v GT, IMR3 v GT, IMR4 v GT, IMR5 v GT, IMR6 v GT, IMR7 v GT, IMR8 v GT	B defined as mean of these eight (8) Kappa values
AIM-NASH	AIM-NASH	AIM-NASH v GT	A defined as this Kappa value

* For IMRs, each slide is read by a minimum of 3 pathologists and maximum of 8 pathologists.

Table 40: Inter-reader Reference Kappa values

Feature	**WK** (6)	**WK** (14)*	WK, PathAI Contributor Network (95% CI)
Steatosis	0.77 (0.69, 0.84)	0.609	0.70 (0.65, 0.75)
Lobular Inflammation	0.46 (0.34, 0.58)	0.328	0.43 (0.38, 0.48)
Hepatocellular Ballooning	0.54 (0.44, 0.65)	0.517	0.42 (0.35, 0.49)
Fibrosis	0.75 (0.67, 0.82)	0.484	0.79 (0.74, 0.85)

* Average WK, no 95% CI were provided in the publication.

Table 41: Intra-reader percent agreement from Davison 2020 (14).

NASH Component	% Agreement*
Steatosis	72.24
Lobular inflammation	55.27
Hepatocellular ballooning	69.92
Fibrosis	71.98

* Pathologist A % agreement used

Determination of Sample Size

Sample size of 600 cases was selected based on Table 42 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 and GT 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 (8), different inter-pathologist Kappas were expected for different components as illustrated in the second column of Table 42. Sample size in the third column of table, is computed based on the simulation described above and inter-pathologist 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 inter-pathologist Kappa - hepatocellular 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 was 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.

Table 42: Sample size from power calculations

Assessment	Expected Linear Kappa for Path-GT & AIM-NASH - GT	Sample Size for 90% Power, Simulation
Steatosis	0.70	360
Lobular inflammation	0.45	540
Hepatocellular ballooning	0.50	600
Fibrosis	0.70	420

4.5.10 AV Results

Study Population

One hundred and sixty-five (165) unique subjects were enrolled in this study from the FALCON 1 clinical trial dataset, 105 unique subjects were enrolled from FALCON 2 clinical trial dataset, and 238 unique subjects were enrolled from the REGENERATE dataset. For each of these trials, not all subjects had more than one time point (baseline and post-treatment) available for enrollment into the study. Overall, 322 samples were enrolled from baseline time point and 283 samples from post-baseline time point. Ultimately, 251 samples from REGENERATE, 217 samples from FALCON 1, and 139 samples were enrolled from FALCON 2 study, with a total of 606 samples from unique biopsies enrolled in AV.

Data Sets Analyzed

Each case included in the accuracy arm of the study has a minimum of 1 AIM-NASH score, a minimum of 2 independent GT pathologist scores, as well as a final score resulting from GT consensus where applicable, , and a set of at least 3 IMR scores per case for each NASH component. All pathologists were blinded to timepoint or any original trial information. Any slide where at least 2 GT pathologists initial reads indicated the biopsy was not adequate for scoring was removed from the analysis for accuracy, repeatability, and reproducibility. Any NASH score component (steatosis, lobular inflammation, hepatocellular ballooning and/or fibrosis) where at least 2 GT pathologists determined the slide to be not evaluable for that component, was removed from the analysis, even if AIM-NASH scores and/or IMR scores were collected for these slides. For any NASH components (steatosis, lobular inflammation, hepatocellular ballooning and/or fibrosis) where the GT pathologist deemed the slide evaluable for the NASH component, but an IMR pathologist deemed it inadequate, that score component was removed from analysis for that IMR pathologist only. In rare cases where AIM-NASH was not able to run on a slide due to little to no evaluable liver tissue being present, the slide was removed from analysis for AIM-NASH only.

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; Table 43).

Table 43: Reason for Missing Final GT Score

Feature	Reason	% (n/N)
Steatosis	Other/reason not given	0.16 (1/607)
	Sample	0.82 (5/607)
	Sample, Scan	0.16 (1/607)
	Sample, Stain	0.33 (2/607)
Inflammation	Other/reason not given	0.33 (2/607)
	Sample	0.99 (6/607)
	Sample, Scan	0.16 (1/607)
	Sample, Stain	0.33 (2/607)
	Stain	0.33 (2/607)
Ballooning	Other/reason not given	0.16 (1/607)
	Sample	0.82 (5/607)
	Sample, Scan	0.16 (1/607)
	Sample, Stain	0.33 (2/607)
Fibrosis	Other/Reason not provided	0.33 (2/607)
	Sample	1.15 (7/607)
	Sample, Scan	0.16 (1/607)
	Sample, Stain	1.15 (7/607)
	Stain	0.99 (6/607)

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.

Demographic and Other Baseline Characteristics

No demographic information for the slides enrolled in the study is available. The dataset represents both screen failed and enrolled NASH clinical trial patient populations, including study subjects who may have regressed or progressed during a clinical trial, and reflects the NASH trial patient population. The dataset also contains variability in sample staining and scanning (including performed by multiple collection/preparation sites and central laboratories). Distribution of slides based on slide level score from GT for accuracy are listed in Table 44. The final score distributions dtid not entirely meet the target distributions identified in the study protocol due to availability of data. However, the study still includes representation at every score component level. Slide distribution by final GT score and sponsor for accuracy are listed in Table 45 Slide distribution for repeatability by AIM-NASH score is listed in Table 46 and reproducibility in Table 47.

Table 44: Slide distribution by final GT for accuracy

Feature	Score	% (n/N)
Steatosis	0	15.58 (93/597)
	1	38.36 (229/597)
	2	30.82 (184/597)
	3	15.24 (91/597)
Lobular Inflammation	0	2.53 (15/593)
	1	57.17 (339/593)
	2	34.91 (207/593)
	3	5.4 (32/593)
Hepatocellular Ballooning	0	12.73 (76/597)
	1	49.41 (295/597)
	2	37.86 (226/597)
Fibrosis	0	3.6 (21/583)
	1	13.38 (78/583)
	2	18.87 (110/583)
	3	33.79 (197/583)
	4	30.36 (177/583)
NAS	0	1.69 (10/593)
	1	5.56 (33/593)
	2	7.76 (46/593)
	3	18.72 (111/593)
	4	22.9 (136/593)
	5	22.4 (133/593)
	6	14.5 (86/593)
	7	5.73 (34/593)
	8	0.67 (4/593)

Table 45: Distribution of cases for accuracy by GT scores and sponsor

Sponsor	Feature	Score	% (n/N)
Falcon 1	Steatosis	0	10.9 (23/211)
		1	42.18 (89/211)
		2	31.75 (67/211)
		3	15.17 (32/211)
	Lobular Inflammation	0	0.48 (1/210)
		1	54.29 (114/210)
		2	38.57 (81/210)
		3	6.67 (14/210)
	Hepatocellular Ballooning	0	7.58 (16/211)
		1	52.61 (111/211)
		2	39.81 (84/211)
	Fibrosis	1	4.27 (9/211)
		2	21.8 (46/211)
		3	55.45 (117/211)
		4	18.48 (39/211)

Sponsor	Feature	Score	% (n/N)
Falcon 2	Steatosis	0	24.64 (34/138)
		1	48.55 (67/138)
		2	21.01 (29/138)
		3	5.8 (8/138)
	Lobular Inflammation	0	2.9 (4/138)
		1	63.04 (87/138)
		2	30.43 (42/138)
		3	3.62 (5/138)
	Hepatocellular Ballooning	0	11.59 (16/138)
		1	52.9 (73/138)
		2	35.51 (49/138)
	Fibrosis	2	2.19 (3/137)
		3	16.06 (22/137)
		4	81.75 (112/137)
Intercept	Steatosis	0	14.52 (36/248)
		1	29.44 (73/248)
		2	35.48 (88/248)
		3	20.56 (51/248)
	Lobular Inflammation	0	4.08 (10/245)
		1	56.33 (138/245)
		2	34.29 (84/245)
		3	5.31 (13/245)
	Hepatocellular Ballooning	0	17.74 (44/248)
		1	44.76 (111/248)
		2	37.5 (93/248)
	Fibrosis	0	8.94 (21/235)
		1	29.36 (69/235)
		2	25.96 (61/235)
		3	24.68 (58/235)
		4	11.06 (26/235)

Table 46: Slide distribution for repeatability based on AIM-NASH

Feature	Score	Day 1 % (n/N)	Day 2 % (n/N)	Day 3 % (n/N)
Steatosis	0	14.48 (21/145)	15.17 (22/145)	15.17 (22/145)
	1	27.59 (40/145)	26.9 (39/145)	27.59 (40/145)
	2	33.79 (49/145)	31.72 (46/145)	31.72 (46/145)
	3	24.14 (35/145)	26.21 (38/145)	25.52 (37/145)
Inflammation	0	14.48 (21/145)	13.79 (20/145)	14.48 (21/145)
	1	35.17 (51/145)	35.17 (51/145)	33.79 (49/145)
	2	44.14 (64/145)	44.14 (64/145)	44.83 (65/145)
	3	6.21 (9/145)	6.9 (10/145)	6.9 (10/145)
Hepatocellular ballooning	0	10.34 (15/145)	10.34, (15/145)	9.66 (14/145)
	1	40.0 (58/145)	39.31 (57/145)	39.31 (57/145)
	2	49.66 (72/145)	50.34 (73/145)	51.03 (74/145)
Fibrosis	0	10.0 (14/140)	10.71 (15/140)	10.71 (15/140)
	1	13.57 (19/140)	12.86 (18/140)	14.29 (20/140)
	2	21.43 (30/140)	20.71 (29/140)	18.57 (26/140)
	3	27.86 (39/140)	29.29 (41/140)	30.0 (42/140)
	4	27.14 (38/140)	26.43 (37/140)	26.43 (37/140)

Table 47: Slide distribution for reproducibility by AIM-NASH

Feature	Score	Accuracy % (n/N)	Repeatability % (n/N)	Reproducibility % (n/N)
Steatosis	0	16.67 (24/144)	15.28 (22/144)	13.89 (20/144)
	1	25 (36/144)	27.08 (39/144)	29.17 (42/144)
	2	31.94 (46/144)	31.25 (45/144)	32.64 (47/144)
	3	26.39, (38/144)	26.39, (38/144)	24.31, (35/144)
Lobular inflammation	0	18.06, (26/144)	13.89, (20/144)	13.19, (19/144)
	1	37.5, (54/144)	35.42, (51/144)	32.64, (47/144)
	2	38.19, (55/144)	43.75, (63/144)	47.92, (69/144)
	3	6.25, (9/144)	6.94, (10/144)	6.25, (9/144)
Hepatocellular ballooning	0	12.5, (18/144)	10.42, (15/144)	9.03, (13/144)
	1	33.33, (48/144)	38.89, (56/144)	35.42, (51/144)
	2	54.17, (78/144)	50.69, (73/144)	55.56, (80/144)
Fibrosis	0	11.51, (16/139)	10.07, (14/139)	8.63, (12/139)
	1	12.95, (18/139)	12.95, (18/139)	17.27, (24/139)
	2	20.86, (29/139)	20.86, (29/139)	16.55, (23/139)
	3	28.78, (40/139)	29.5, (41/139)	30.22, (42/139)
	4	25.9, (36/139)	26.62, (37/139)	27.34, (38/139)

Accuracy Results

Primary Endpoints for Accuracy

Evaluation for non-inferior accuracy of AIM-NASH to IMR was assessed by comparing the WK of IMR with GT to the WK of AIM-NASH with GT (Figure 21 and Figure 22 and Table 48). The difference in WKs of AIM-NASH with GT, and WKs of IMR with GT for hepatocellular ballooning was 0.168 (95% CI of (0.098, 0.252), NI p<0.0001), indicating superiority to IMRs with p<0.0001. The difference in WKs for steatosis, lobular inflammation, and fibrosis were -0.045 (95% CI of (-0.095, 0.006), NI p=0.016), 0.01 (95% CI of (-0.063, 0.104), NI p=0.0005), and 0.024 (95% CI of (-0.023, 0.088), NI p<0.0001) respectively, demonstrating non-inferiority to IMRs. Steatosis, lobular inflammation, and fibrosis did not show superiority to IMRs. Overall, WKs compared to GT for both manual and AIM-NASH reads were similar to the range of reported inter-pathologist kappas demonstrated by CRN and other experts (Table 40).

Figure 21: Accuracy results for each NASH component

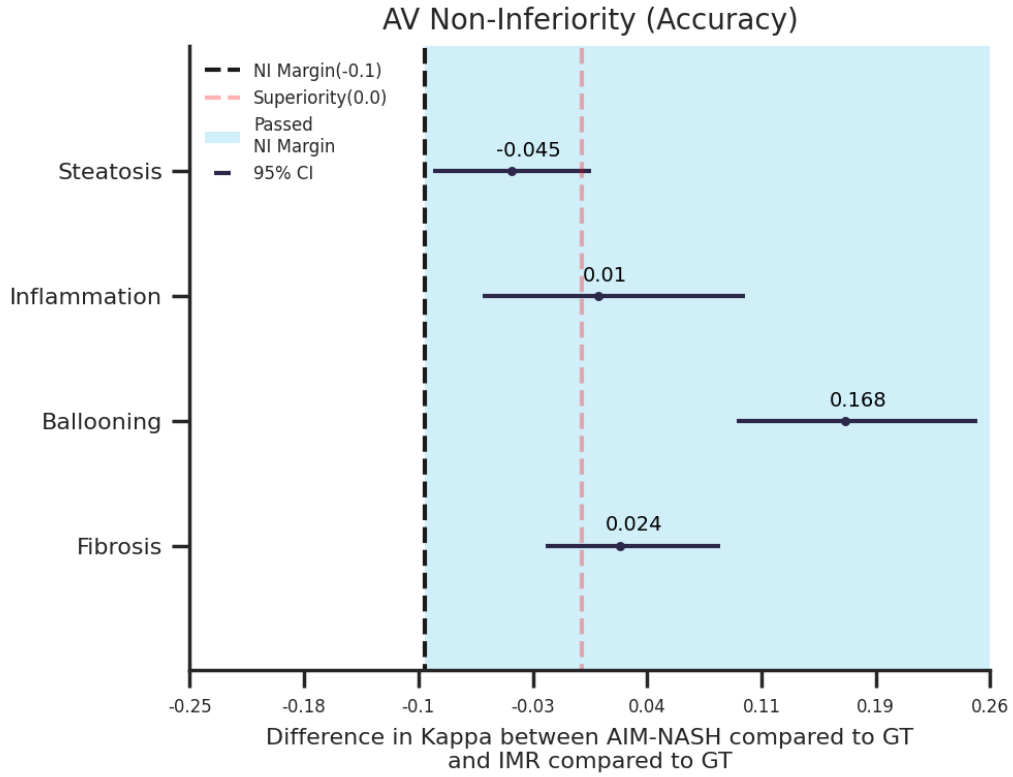


Figure 22: Accuracy concordance comparison of NASH component scores

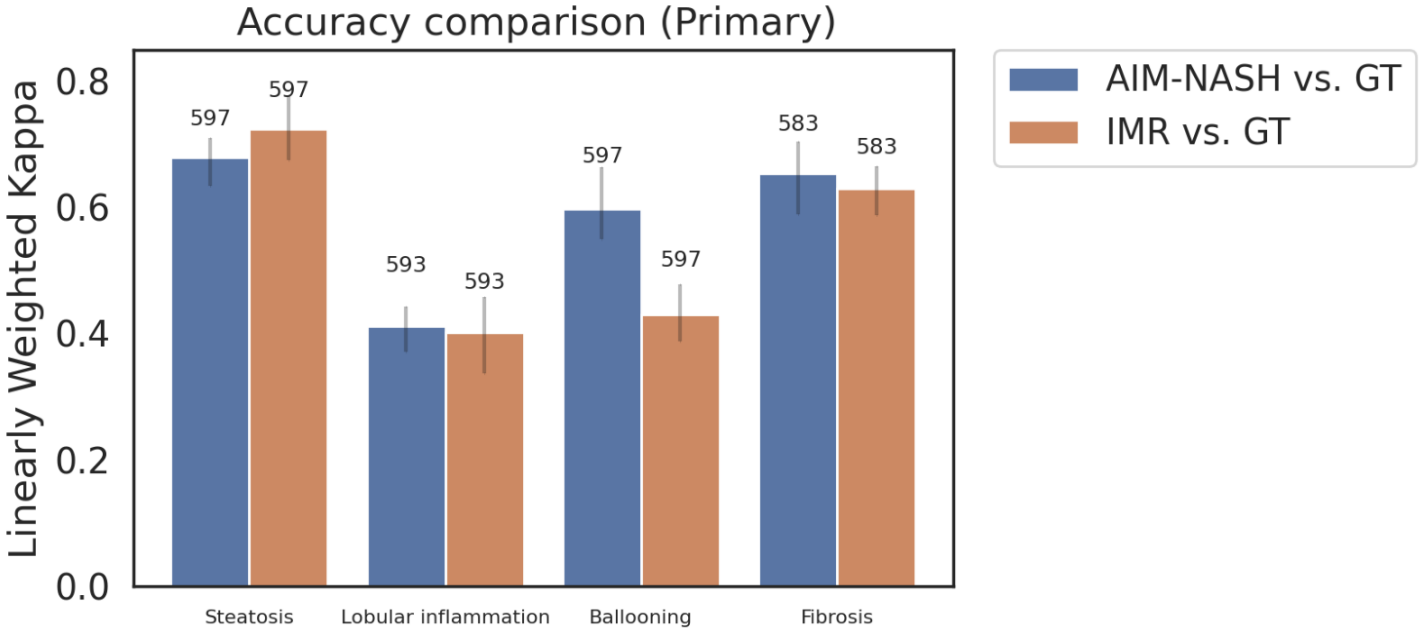


Table 48: Primary endpoint results for accuracy

Feature	Modality Comparison	N	WK (95% CI)	Difference (95% CI)	p-value for NI	p-value for Superiority
Steatosis	AIM-NASH vs GT	597	0.679 (0.634, 0.711)	-0.045 (-0.095, 0.006)	0.016	0.954
	IMR vs GT	597	0.724 (0.683, 0.755)			
Lobular inflammation	AIM-NASH vs GT	593	0.412 (0.365, 0.479)	0.01 (-0.063, 0.104)	0.0005	0.343
	IMR vs GT	593	0.402 (0.337, 0.452)			
Hepatocellular ballooning	AIM-NASH vs GT	597	0.597 (0.548, 0.651)	0.168 (0.098, 0.252)	<0.0001	<0.0001
	IMR vs GT	597	0.430 (0.365, 0.486)			
Fibrosis	AIM-NASH vs GT	583	0.654 (0.612, 0.702)	0.024 (-0.023, 0.088)	<0.0001	0.1325
	IMR vs GT	583	0.630 (0.587, 0.665)			

Secondary Endpoints for Accuracy

The study was not powered for any of the secondary endpoints. For secondary endpoint analysis, WKs for AIM-NASH and GT vs IMR and GT were calculated for NAS aggregate scores including F0&1 vs other, F4 vs other and NAS >= 4 vs NAS<4. WK for AIM-NASH and GT was higher compared to WK for IMR and GT across all categories of scores relevant in commonly used trial inclusion/exclusion criteria (Figure 23, Table 49 and Table 50). The difference in WK for F0&1 vs other was 0.088 (95% CI of -068, 0.187) and the difference for F4 vs other was 0.076 (95% CI of -0.037,0.179). The difference for overall NAS $\geq$4 was 0.118 (95% CI of 0.026, 0.194).

Figure 23: Accuracy agreement comparison within common trial inclusion criteria score groups

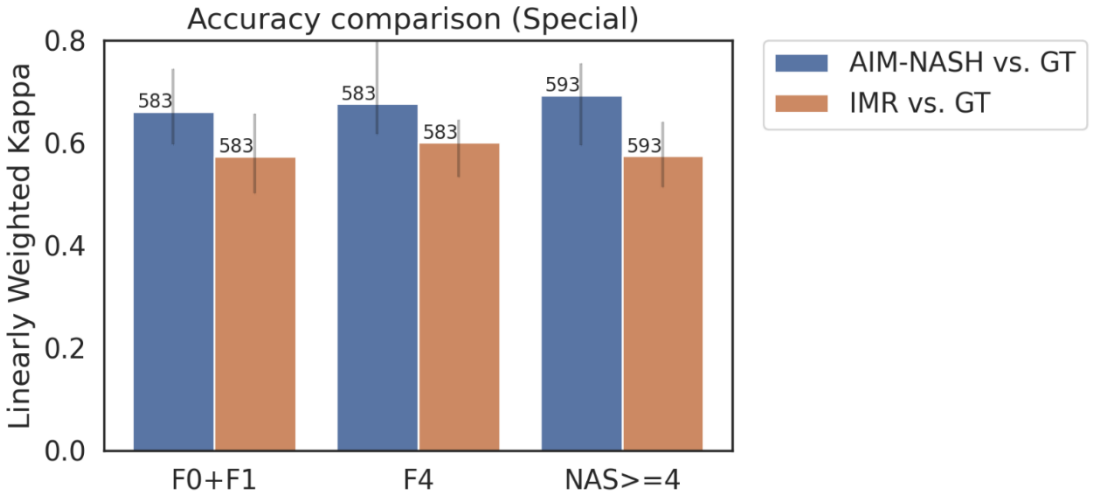


Table 49: WKs for NASH aggregate scores for accuracy

Aggregate Score	Modality	N	WK (95% CI)	Difference (95% CI)
F0&F1 vs other	AIM-NASH vs GT	583	0.661 (0.597, 0.744)	0.088 (-0.068, 0.187)
F0&F1 vs other	IMR vs GT	583	0.573 (0.513, 0.705)	
F4 vs other	AIM-NASH vs GT	583	0.676 (0.579, 0.738)	0.076 (-0.037, 0.179)
F4 vs other	IMR vs GT	583	0.600 (0.529, 0.684)	
NAS ≥4 vs. <4	AIM-NASH vs GT	593	0.692 (0.625, 0.737)	0.118 (0.026, 0.194)
NAS ≥4 vs. <4	IMR vs GT	593	0.574 (0.513, 0.641)	

Table 50: WKs for NAS aggregate score F2&3 vs other, NAS >4 with 1> in each component and NASH resolution

Aggregate Score	Modality	N	WK (95% CI)	Difference (95% CI)
NAS ≥4 and ≥1 for each component vs. Other	AIM-NASH vs GT	593	0.701 (0.642, 0.749)	0.165 (0.082, 0.239)
	IMR vs GT	593	0.536 (0.48, 0.6)	
F2&F3 vs other	AIM-NASH vs GT	583	0.541 (0.461, 0.621)	0.069 (-0.028, 0.183)
	IMR vs GT	583	0.472 (0.398, 0.555)	
NASH resolution	AIM-NASH vs GT	593	0.595 (0.517, 0.679)	0.276 (0.17, 0.38)
	IMR vs GT	593	0.319 (0.26, 0.393)	

Secondary analyses were also performed for each NASH component by time point (baseline and post-baseline). For lobular inflammation and fibrosis, AIM-NASH concordance with GT was non-inferior to IMR concordance with GT for both baseline and post-baseline timepoints (Figure 24, Table 51). The difference between AIM-NASH and IMR with GT was 0.03 (95% CI of -0.066, 0.148) at baseline and 0.038 (95% CI of -0.05, 0.135) at post-baseline for lobular inflammation, and -0.004 (95% CI of -0.083, 0.091) at baseline and 0.049 (95% CI of -0.045, 0.151) at post-baseline for fibrosis. Hepatocellular ballooning concordance between AIM-NASH and GT was significantly higher to manual reads for both baseline and post-baseline timepoints, with the difference of 0.133 (95% CI of 0.033, 0.236) at baseline and 0.197 (95% CI of 0.067, 0.339) post-baseline. For steatosis the difference

between AIM-NASH and IMR with GT was -0.051 (95% CI of -0.134, 0.024) at baseline and -0.037 (95% CI of -0.109, 0.041) at post-baseline.

Figure 24: WKs for NASH components per time point

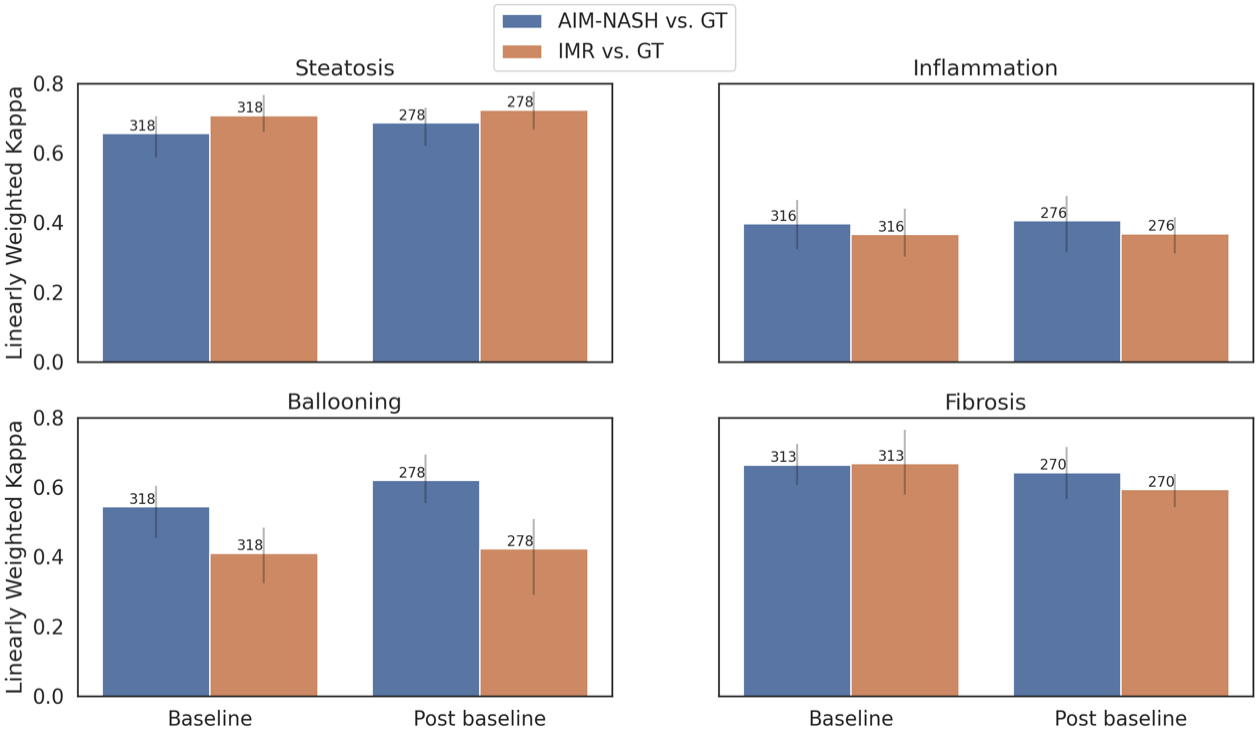


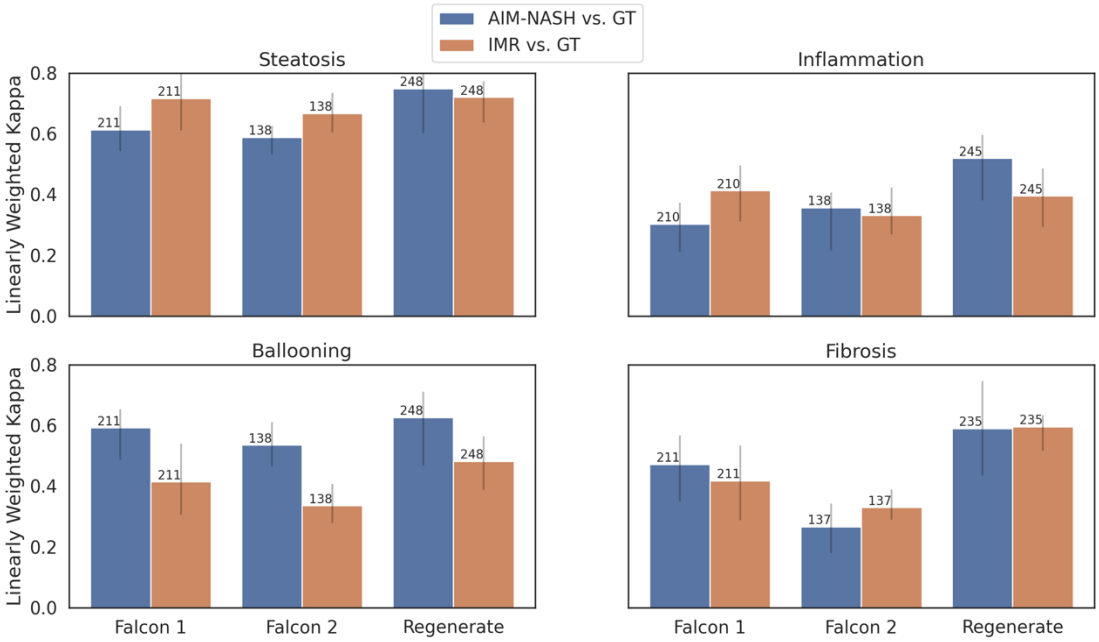
Table 51: WKs for NASH components per time point for accuracy

Feature	Visit	Comparison	N	WK (95% CI)	Difference (95% CI)
Steatosis	Baseline	AIM-NASH vs GT	318	0.657 (0.589, 0.704)	-0.051 (-0.134, 0.024)
		IMR vs GT	318	0.708 (0.643, 0.75)	
	Post baseline	AIM-NASH vs GT	278	0.687 (0.64, 0.745)	-0.037 (-0.109, 0.041)
		IMR vs GT	278	0.724 (0.668, 0.776)	
Lobular inflammation	Baseline	AIM-NASH vs GT	316	0.397 (0.324, 0.464)	0.03 (-0.066, 0.148)
		IMR vs GT	316	0.366 (0.277, 0.436)	
	Post baseline	AIM-NASH vs GT	276	0.406 (0.342, 0.479)	0.038 (-0.05, 0.135)
		IMR vs GT	276	0.368 (0.311, 0.415)	
Hepatocellular ballooning	Baseline	AIM-NASH vs GT	318	0.544 (0.455, 0.602)	0.133 (0.033, 0.236)
		IMR vs GT	318	0.411 (0.345, 0.484)	
	Post baseline	AIM-NASH vs GT	278	0.620 (0.534, 0.693)	0.197 (0.067, 0.339)
		IMR vs GT	278	0.423 (0.29, 0.508)	
Fibrosis	Baseline	AIM-NASH vs GT	313	0.664 (0.606, 0.724)	-0.004 (-0.083, 0.091)
		IMR vs GT	313	0.668 (0.592, 0.74)	
	Post baseline	AIM-NASH vs GT	270	0.642 (0.554, 0.739)	0.049 (-0.045, 0.151)
		IMR vs GT	270	0.593 (0.543, 0.637)	

Predefined Exploratory Endpoints for Accuracy

Predefined exploratory analysis per trial of origin for each NASH component is shown in Figure 25 and Table 52. The study was not powered for any exploratory endpoints. For hepatocellular ballooning the WKs for AIM-NASH and GT were higher than the WKs for IMR and GT for all 3 clinical trials (FALCON 1, FALCON 2 and REGENERATE). Some differences were observed between AIM vs GT and IMR vs GT across clinical trials for steatosis, lobular inflammation, and fibrosis.

Figure 25: WKs per NASH component per trial of origin for accuracy



Analysis of agreement of AIM-NASH with GT compared to agreement of IMR with GT was performed for each score level for all NASH components. Agreements for score levels within steatosis, inflammation, and fibrosis were largely similar, with overlapping confidence intervals for IMRs vs GT and AIM-NASH vs. GT, with some average WKs being higher for AIM-NASH and some being higher for IMRs, within each component. As was demonstrated with time point and trial subsets, hepatocellular ballooning accuracy was significantly higher with AIM-NASH at each score (Figure 26 and Table 53).

Table 52: WKs for NASH components per trial of origin for accuracy

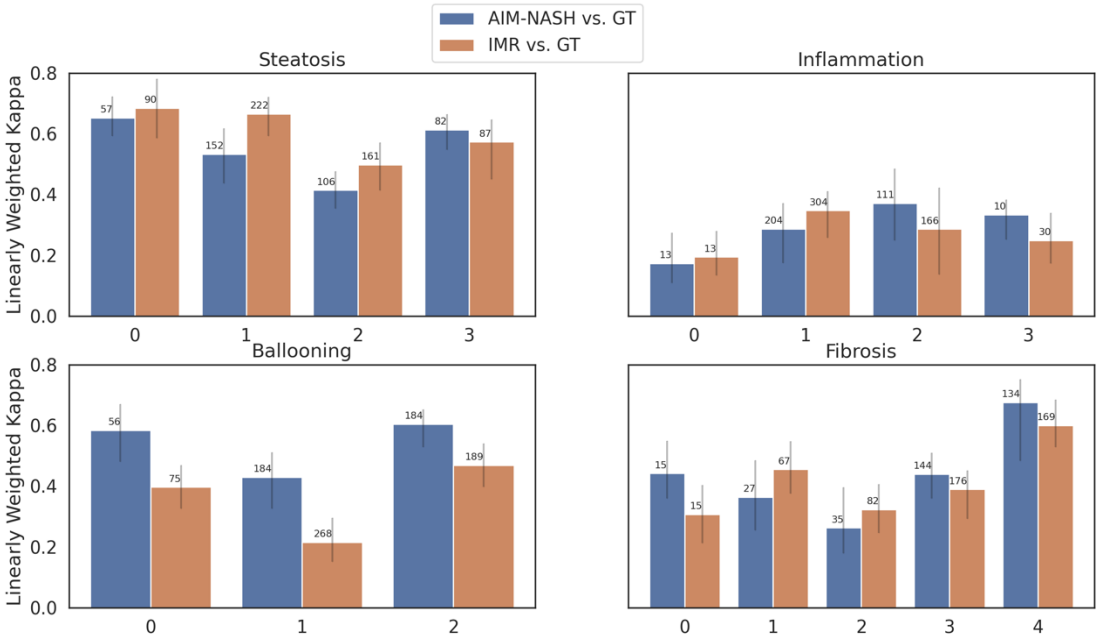
Component	Trial	Modality	N	WK (95% CI)	Difference (95% CI)
Steatosis	Falcon 1	AIM vs GT	211	0.613 (0.543,0.69)	-0.104 (-0.178, -0.022)
		IM vs GT	211	0.717 (0.659, 0.753)	
	Falcon 2	AIM vs GT	138	0.588 (0.443,0.658)	-0.079 (-0.243, 0.046)
		IM vs GT	138	0.667 (0.561, 0.763)	
	Regenerate	AIM vs GT	248	0.749 (0.687, 0.815)	0.028 (-0.046, 0.136)
		IM vs GT	248	0.721 (0.638, 0.772)	
Lobular inflammation	Falcon 1	AIM vs GT	210	0.303 (0.213, 0.372)	-0.11 (-0.203,0.062)
		IM vs GT	210	0.413 (0.274, 0.462)	
	Falcon 2	AIM vs GT	138	0.357 (0.218, 0.433)	0.025 (-0.136,0.158)
		IM vs GT	138	0.332 (0.23, 0.414)	
	Regenerate	AIM vs GT	245	0.519 (0.456, 0.61)	0.124 (0.019,0.264)
		IM vs GT	245	0.395 (0.294, 0.484)	
Hepatocellular ballooning	Falcon 1	AIM vs GT	211	0.592 (0.487, 0.653)	0.178 (0.055,0.277)
		IM vs GT	211	0.414 (0.343, 0.489)	
	Falcon 2	AIM vs GT	138	0.535 (0.376, 0.62)	0.2 (0.009,0.339)
		IM vs GT	138	0.336 (0.228, 0.46)	
	Regenerate	AIM vs GT	248	0.626 (0.569, 0.697)	0.144 (0.059,0.272)
		IM vs GT	248	0.482 (0.388, 0.563)	
Fibrosis	Falcon 1	AIM vs GT	211	0.472 (0.351, 0.566)	0.054 (-0.093,0.175)
		IM vs GT	211	0.418 (0.332, 0.493)	
	Falcon 2	AIM vs GT	137	0.267 (0.113, 0.423)	-0.063 (-0.253,0.148)
		IM vs GT	137	0.330 (0.2, 0.444)	
	Regenerate	AIM vs GT	235	0.589 (0.549, 0.648)	-0.006 (-0.06,0.098)
		IM vs GT	235	0.596 (0.517, 0.633)	

Table 53: WKs for NASH components per score for accuracy

Feature	Score	Modality	N	WK (95% CI)	Difference (95% CI)
Steatosis	0	AIM-NASH vs GT	57	0.653 (0.592, 0.722)	-0.032 (-0.139, 0.098)
		IMR vs GT	90	0.685 (0.588, 0.769)	
	1	AIM-NASH vs GT	152	0.533 (0.473, 0.595)	-0.133 (-0.218, -0.053)
		IMR vs GT	222	0.666 (0.6, 0.717)	
	2	AIM-NASH vs GT	106	0.415 (0.314, 0.51)	-0.083 (-0.188, 0.026)
		IMR vs GT	161	0.498 (0.424, 0.552)	
	3	AIM-NASH vs GT	82	0.613 (0.528, 0.686)	0.04 (-0.07, 0.185)
		IMR vs GT	87	0.573 (0.45, 0.646)	
Lobular inflammation	0	AIM-NASH vs GT	13	0.173 (0.109, 0.273)	-0.022 (-0.125, 0.126)
		IMR vs GT	13	0.194 (0.082, 0.28)	

Feature	Score	Modality	N	WK (95% CI)	Difference (95% CI)
	1	AIM-NASH vs GT	204	0.287 (0.164, 0.4)	-0.061 (-0.178, 0.068)
		IMR vs GT	304	0.348 (0.266, 0.397)	
	2	AIM-NASH vs GT	111	0.372 (0.31, 0.456)	0.086 (-0.01, 0.204)
		IMR vs GT	166	0.286 (0.196, 0.349)	
	3	AIM-NASH vs GT	10	0.334 (0.184, 0.469)	0.085 (-0.078, 0.238)
		IMR vs GT	30	0.249 (0.173, 0.339)	
Hepatocellular ballooning	0	AIM-NASH vs GT	56	0.584 (0.481, 0.669)	0.187 (0.075, 0.335)
		IMR vs GT	75	0.397 (0.294, 0.479)	
	1	AIM-NASH vs GT	184	0.429 (0.354, 0.477)	0.214 (0.118, 0.299)
		IMR vs GT	268	0.215 (0.145, 0.287)	
	2	AIM-NASH vs GT	184	0.604 (0.541, 0.684)	0.136 (0.045, 0.229)
		IMR vs GT	189	0.469 (0.397, 0.54)	
Fibrosis	0	AIM-NASH vs GT	15	0.442 (0.36, 0.549)	0.134 (-0.026, 0.307)
		IMR vs GT	15	0.307 (0.198, 0.428)	
	1	AIM-NASH vs GT	27	0.364 (0.28, 0.496)	-0.091 (-0.2, 0.069)
		IMR vs GT	67	0.455 (0.375, 0.524)	
	2	AIM-NASH vs GT	35	0.263 (0.069, 0.337)	-0.06 (-0.231, 0.061)
		IMR vs GT	82	0.323 (0.227, 0.419)	
	3	AIM-NASH vs GT	144	0.440 (0.36, 0.531)	0.05 (-0.067, 0.169)
		IMR vs GT	176	0.390 (0.312, 0.472)	
	4	AIM-NASH vs GT	134	0.676 (0.579, 0.738)	0.076 (-0.037, 0.179)
		IMR vs GT	169	0.600 (0.529, 0.684)	

Figure 26: WKs for NASH components by score for accuracy



Post-hoc Exploratory Analysis Endpoints for Accuracy

Post-hoc exploratory analysis included additional NAS $\geq$ 4 with a score of at least 1 for each component (secondary endpoint only included NAS $\geq$ 4 vs NAS <4), NAS component score F2&F3 vs other (commonly included for non-cirrhotic trial enrollment criteria) and NASH resolution (defined as ballooning score of 0, lobular inflammation score of 0 or 1, and any value for steatosis; fibrosis scores were not included). For all three of these analyses (Figure 27 and Table 54), AIM-NASH showed significantly better or non-inferior agreement with GT than IMRs. For F2&3 (commonly enrolled in non-cirrhotic trials) the difference in WK was 0.069 (95% CI of -0.028, 0.183). The difference for overall NAS $\geq$ 4, with a score of $\geq$1 for each component was 0.165 (95% CI of 0.082, 0.239), and the difference for NASH resolution (based on H&E scores) was 0.276 (95% CI of 0.17, 0.38), with AIM-NASH agreement with GT being strikingly higher than for IMR and GT.

Figure 27: WKs for NAS aggregate score F2&3 vs other, NAS >4 with 1> in each component and NASH resolution.

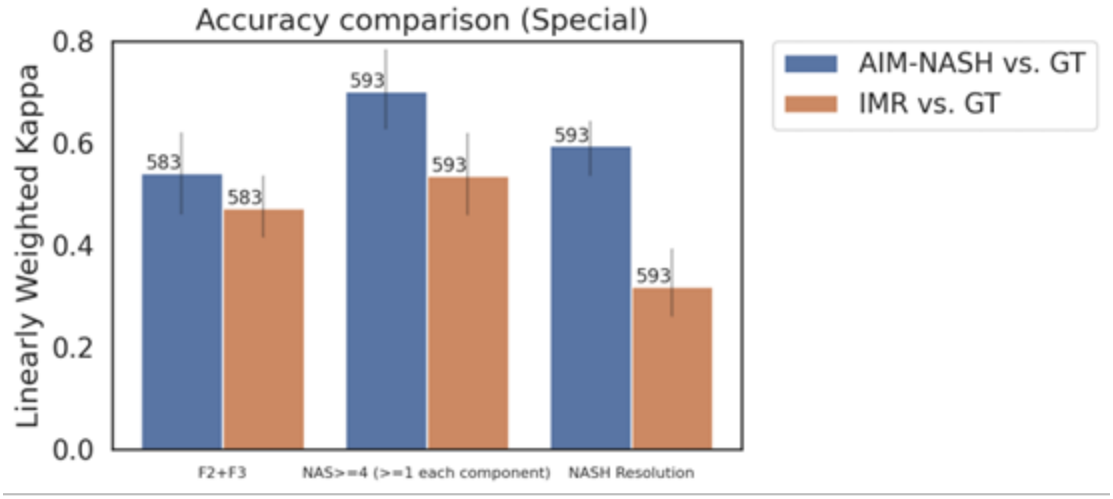


Table 54: WKs for NAS aggregate score F2&3 vs other, NAS >4 with 1> in each component and NASH resolution

Aggregate Score	Modality	N	WK (95% CI)	Difference (95% CI)
NAS ≥4 and ≥1 for each component vs. Other	AIM-NASH vs GT	593	0.701 (0.642, 0.749)	0.165 (0.082, 0.239)
	IMR vs GT	593	0.536 (0.48, 0.6)	
F2&F3 vs other	AIM-NASH vs GT	583	0.541 (0.461, 0.621)	0.069 (-0.028, 0.183)
	IMR vs GT	583	0.472 (0.398, 0.555)	
NASH resolution (without fibrosis)	AIM-NASH vs GT	593	0.595 (0.517, 0.679)	0.276 (0.17, 0.38)
	IMR vs GT	593	0.319 (0.26, 0.393)	

Post-hoc accuracy analysis was repeated for all 606 cases after non-liver tissue was excluded from the WSIs and AIM-NASH was re-run (Table 55). Eighteen (18) percent of cases had non-liver tissue present. After non-liver tissue exclusion, the accuracy results did not change significantly and differences in WKs for AIM-NASH and GT vs IMR and GT remained similar for all NASH components.

Table 55: WKs for AIM-NASH accuracy after non-liver tissue exclusion

Feature	Modality	N	WK (95% CI)	Difference (95% CI)	p-value for NI	p-value for superiority
Steatosis	AIM-NASH vs GT	597	0.683 (0.642, 0.722)	-0.041 (-0.094, 0.016)	0.0145	0.921
	IMR vs GT	597	0.724 (0.683, 0.755)			
Lobular inflammation	AIM-NASH vs GT	593	0.414 (0.361, 0.468)	0.012 (-0.063, 0.098)	0.0025	0.362
	IMR vs GT	593	0.402 (0.337, 0.452)			
Hepatocellular ballooning	AIM-NASH vs GT	597	0.597 (0.54, 0.652)	0.168 (0.088, 0.253)	<0.0001	<0.0001
	IMR vs GT	597	0.430 (0.365, 0.486)			
Fibrosis	AIM-NASH vs GT	583	0.649 (0.61, 0.687)	0.02 (-0.033, 0.081)	<0.0001	0.2015
	IMR vs GT	583	0.630 (0.587, 0.665)			

Results for Repeatability

Primary Endpoint for Repeatability

Repeatability endpoint for this study included scans from the same glass slides repeated over 3 non-consecutive days. The AIM-NASH was then deployed on each WSI and the agreement per histologic component was evaluated across WSIs. Mean agreement rates between the AIM-NASH scoring on the 3 separate WSIs for steatosis was 0.931 (95% CI of (0.894, 0.963), p<0.0001), lobular inflammation was 0.963 (95% CI of (0.937, 0.986), p<0.0001), hepatocellular ballooning was 0.958 (95% CI of (0.931, 0.982), p<0.0001) and fibrosis was 0.926 (95% CI of (0.891, 0.96), p<0.001), (Figure 28, Table 56). This demonstrates superior performance when comparing to a performance goal of 85% as well as relevant published manual intra-pathologist read agreements (steatosis 0.722, lobular inflammation 0.553, hepatocellular ballooning 0.699 and fibrosis 0.720) described in the literature (14).

Figure 28: Repeatability mean percent agreement rate by NASH component

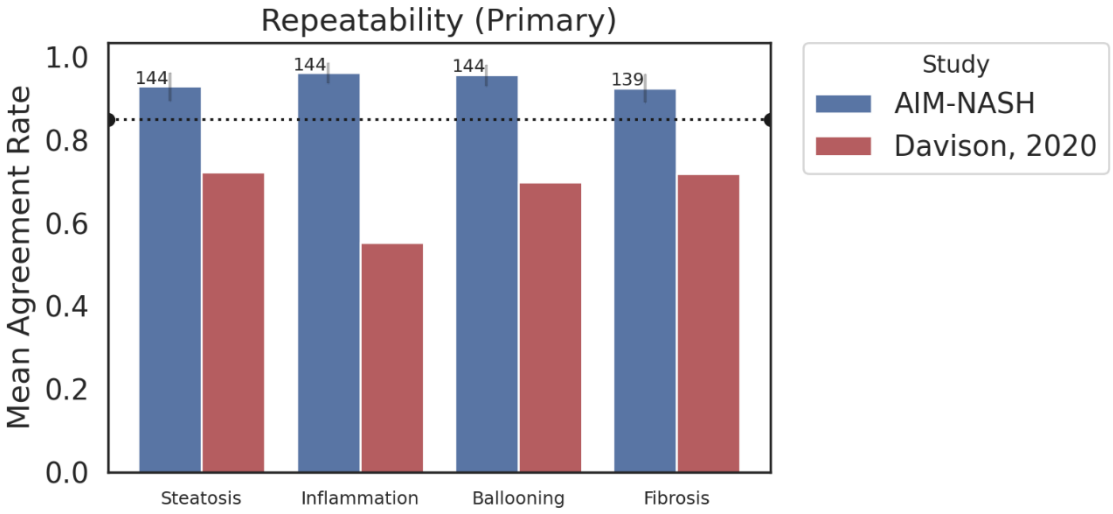


Table 56: Mean agreement rates between the AIM-NASH scoring on the 3 separate WSIs for all NASH components

Feature	N*	Agreement (95% CI)	p-value
Steatosis	597	0.931, (0.894, 0.963)	<0.0001
Lobular inflammation	593	0.963, (0.937, 0.986)	<0.0001
Hepatocellular ballooning	597	0.958, (0.931, 0.982)	<0.0001
Fibrosis	583	0.926, (0.891, 0.96)	<0.0001

* this indicates the number of WSIs in each pairwise agreement

Secondary Endpoints for Repeatability

Mean agreement of AIM-NASH scoring between inter-day timepoints were assessed across baseline and post-baseline timepoints (including placebo and treatment groups) of the FALCON 1, FALCON 2 and REGENERATE datasets (Figure 29, Table 57). The mean agreement of all time points was significantly greater than 0.85.

Figure 29: Mean agreement of AIM-NASH scoring per time-points for repeatability

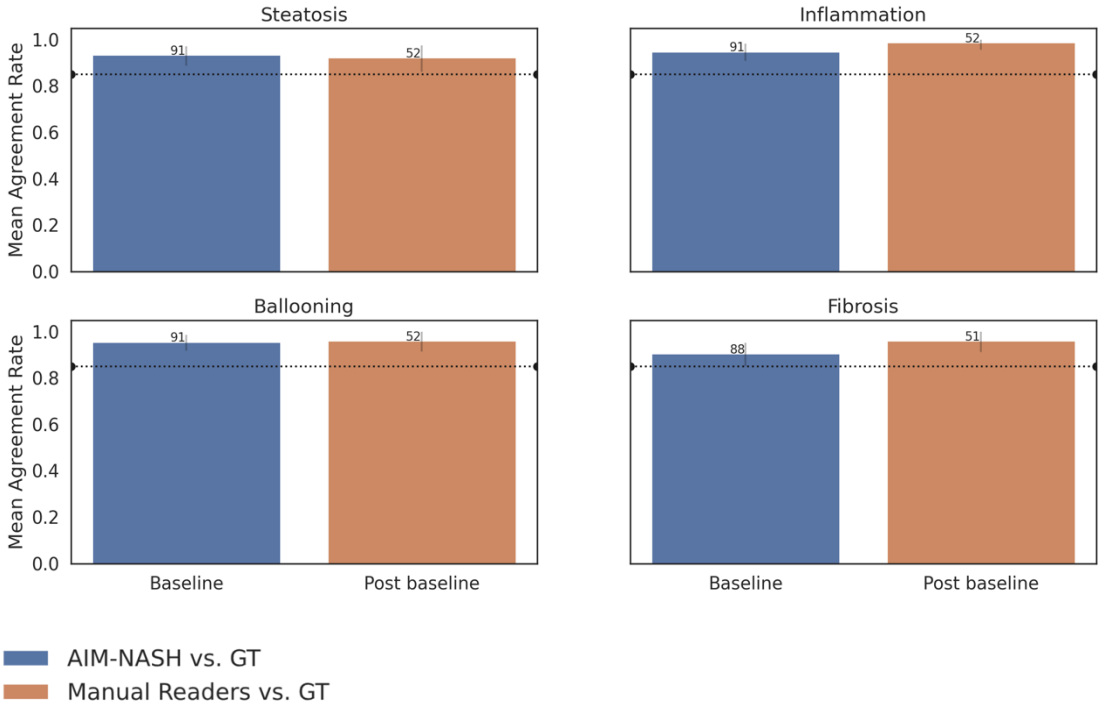


Table 57: Mean agreement rate by time points for repeatability

Feature	Visit	N	Mean Agreement Rate (95% CI)
Steatosis	Baseline	91	0.934, (0.889, 0.972)
	Post baseline	52	0.923, (0.862, 0.975)
Lobular inflammation	Baseline	91	0.949, (0.91, 0.983)
	Post baseline	52	0.987, (0.957, 1)
Hepatocellular ballooning	Baseline	91	0.956, (0.92, 0.986)
	Post baseline	52	0.962, (0.915, 1)
Fibrosis	Baseline	88	0.905, (0.852, 0.952)
	Post baseline	51	0.961, (0.914, 1)

Predefined Exploratory Endpoints for Repeatability

Exploratory analysis assessed each NASH component score for mean percent agreement. Observed concordance between AIM-NASH scoring at inter-day time points was higher than 85% for all score levels for steatosis, lobular inflammation, and hepatocellular ballooning and fibrosis (Figure 30, Table 58), except for fibrosis score 1 (mean agreement rate of 0.808 (95% CI of 0.679, 0.912), however this agreement rate is still higher than reported intra-reader agreement for fibrosis overall (0.72) (14).

Figure 30: Mean agreement rate by score for repeatability

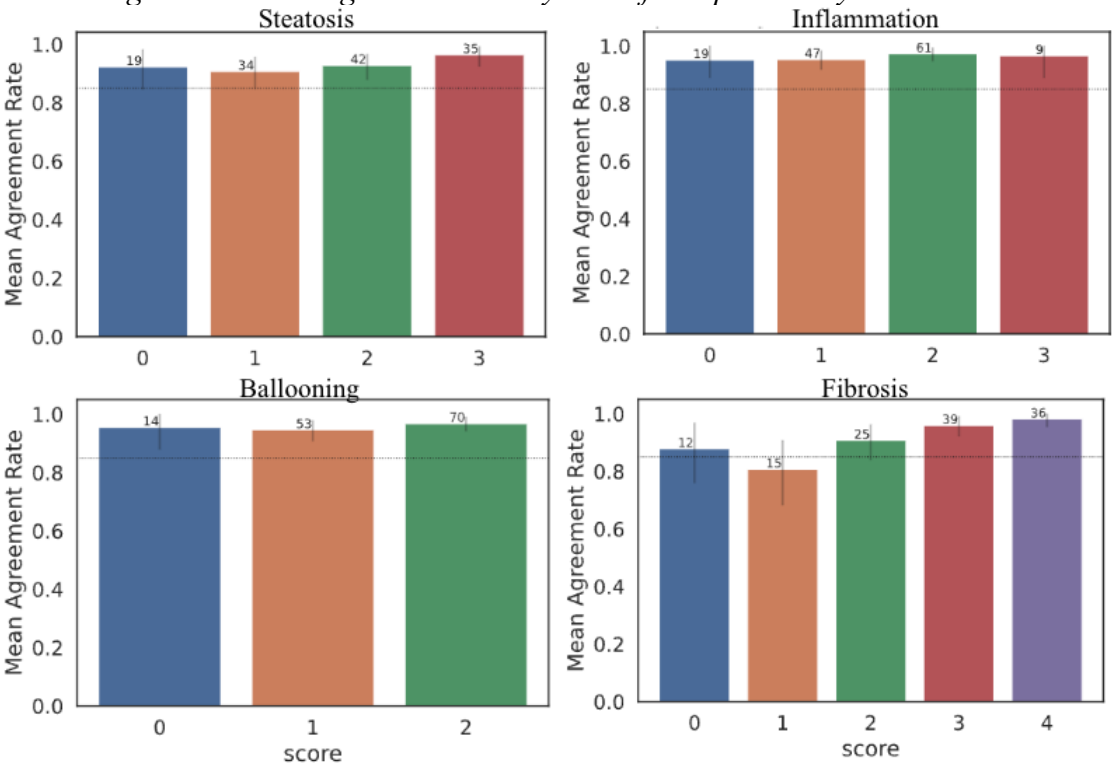


Table 58: Mean agreement rate by score for repeatability

Feature	Score	N	Mean Agreement (95% CI)
Steatosis	0	19	0.923, (0.846, 0.983)
	1	34	0.908, (0.846, 0.956)
	2	42	0.928, (0.88, 0.967)
	3	35	0.965, (0.925, 0.992)
Lobular inflammation	0	19	0.952, (0.889, 1)
	1	47	0.954, (0.918, 0.984)
	2	61	0.974, (0.948, 0.995)
	3	9	0.967, (0.889, 1)
Hepatocellular ballooning	0	14	0.956, (0.88, 1)
	1	53	0.947, (0.908, 0.978)
	2	70	0.968, (0.942, 0.99)
Fibrosis	0	12	0.879, (0.76, 0.969)
	1	15	0.808, (0.682, 0.907)
	2	25	0.908, (0.839, 0.962)
	3	39	0.960, (0.922, 0.991)
	4	36	0.982, (0.954, 1)

Each of the 3 trial subsets also demonstrated significantly above 85% agreement for all histologic components (Figure 31 and Table 59), except for steatosis for FALCON 1 0.908 (95% CI of 0.837,0.966) and fibrosis for REGENERATE 0.869 (95% CI, 0.797, 0.933).

Figure 31: Mean agreement rate by trial of origin for repeatability.

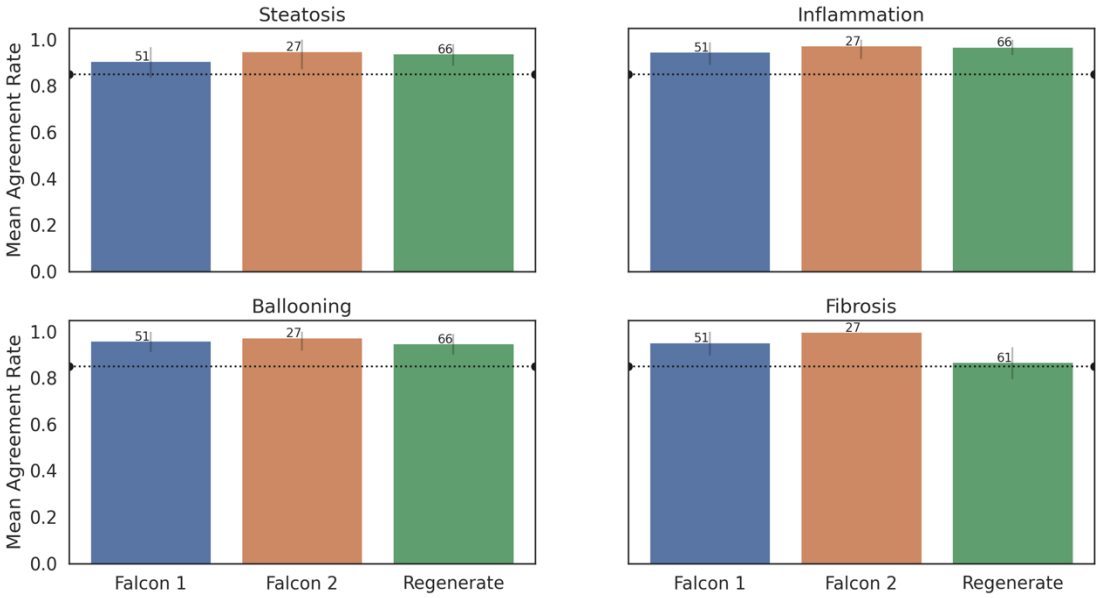


Table 59: Mean agreement rate by trial of origin for repeatability

Feature	Trial	N	Agreement Rate (95% CI)
Steatosis	Falcon 1	51	0.908 (0.837, 0.966)
	Falcon 2	27	0.951 (0.873, 1)
	Regenerate	66	0.939 (0.889, 0.981)
Lobular inflammation	Falcon 1	51	0.948 (0.892, 0.988)
	Falcon 2	27	0.975 (0.917, 1)
	Regenerate	66	0.970 (0.934, 1)
Hepatocellular ballooning	Falcon 1	51	0.961 (0.913, 1)
	Falcon 2	27	0.975 (0.92, 1)
	Regenerate	66	0.949 (0.902, 0.989)
Fibrosis	Falcon 1	51	0.954 (0.898, 1)
	Falcon 2	27	1.000 (0.875, 1)*
	Regenerate	61	0.869 (0.797, 0.933)

* Since the agreement rate is 1, the 95% CI is based on Wilson score confidence interval

Post-hoc Exploratory Endpoints for Repeatability

In order to explicitly demonstrate that the AIM-NASH algorithm is 100% repeatable when it is run multiple times on the same WSI, AIM-NASH was deployed 5 separate times on each of the 150 repeatability cases (all from day 2) and achieved perfect agreement across all histologic components (Table 60).

Table 60: Mean agreement rate for AIM-NASH when deployed on the same WSI 5 times

Feature	N	Agreement Rate (95% CI)*
Steatosis	150	1 (0.975, 1)
Lobular inflammation	150	1 (0.975, 1)
Hepatocellular ballooning	150	1 (0.975, 1)
Fibrosis	150	1 (0.975, 1)

* Since the agreement rate is 1, the 95% CI is based on Wilson score confidence interval

Results for Reproducibility

Primary Endpoints for Reproducibility

To evaluate the reproducibility of AIM-NASH, the algorithm was run on WSIs obtained from the 150 cases, scanned at three different external sites using the Aperio AT2 scanner with different operators. The mean agreement rate for hepatocellular ballooning was 0.912 (95% CI of 0.872, 0.949, p=0.02), meeting the acceptance criteria for reproducibility. The mean agreement rates for steatosis, lobular inflammation, and fibrosis were approximately 85% but the CIs fell slightly below (steatosis 0.856 (95% CI of 0.808, 0.9, p=0.389), lobular inflammation 0.847 (95% CI of 0.8, 0.891, p=0.532), and fibrosis 0.868 (95% CI of 0.823, 0.911, p=0.207) (Figure 32, Table 61). However, AIM-NASH reproducibility across the three external laboratories, utilizing different operators and different Aperio AT2 scanners, was still higher for all NASH components than published literature across expert NASH pathologists (0.633 for steatosis, 0.604 for lobular inflammation, 0.625 for hepatocellular ballooning and 0.509 for fibrosis; (14).

Figure 32: Mean agreement rate by NASH component for reproducibility compared to published inter-reader agreement

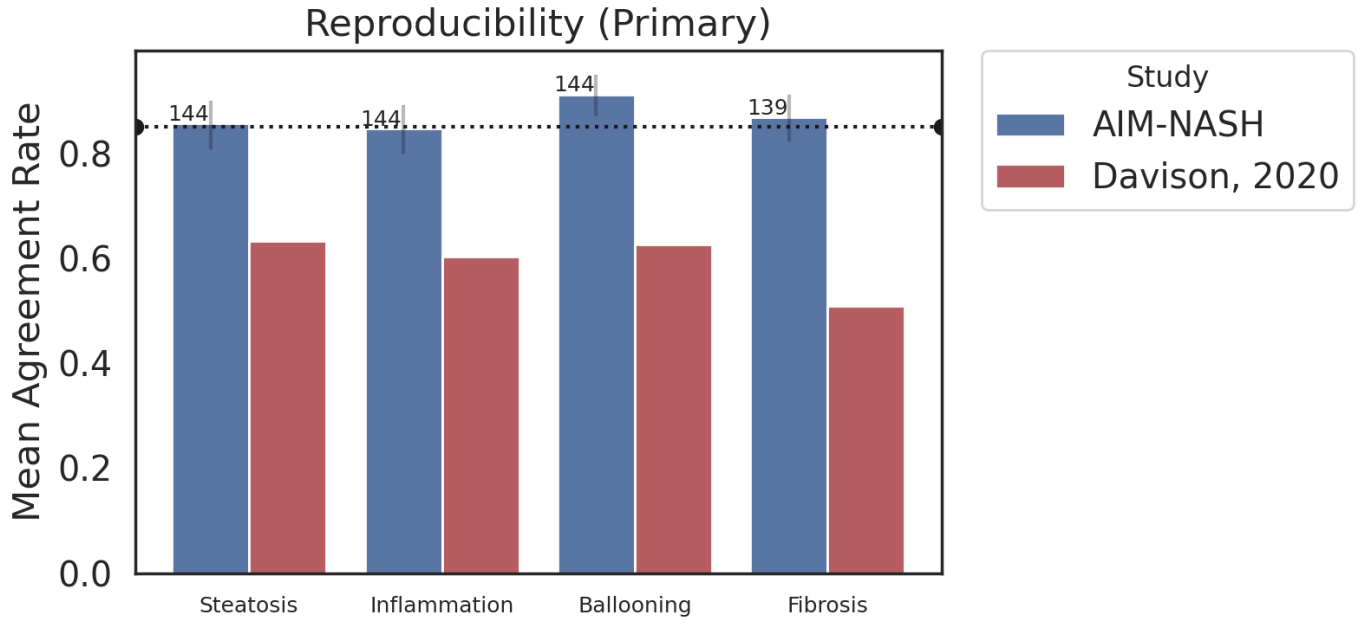


Table 61: Mean agreement rate by NASH component for reproducibility

Feature	N*	Agreement, 95% CI	p-value
Steatosis	144	0.856, (0.808, 0.9)	0.389
Lobular inflammation	144	0.847, (0.8, 0.891)	0.532
Hepatocellular ballooning	144	0.912, (0.872, 0.949)	0.002
Fibrosis	139	0.868, (0.823, 0.911)	0.207

* this indicates the number of WSIs in each pairwise agreement

Secondary Analysis Endpoints for Reproducibility

Secondary endpoint analyses of agreement of AIM-NASH scoring from slides scanned on multiple days on 3 different Aperio AT2 scanners, were performed per time points (baseline and post baseline datasets (Figure 33, Table 62). AIM-NASH was superior to 85% for ballooning scores at the baseline timepoints, with a rate of 0.912 (95% CI of 0.865, 0.957). The observed average agreement was near 85% or greater for all NASH components and time points except for steatosis post baseline, (0.808) and lobular inflammation baseline (0.824).

Figure 33: Mean agreement by time point for reproducibility

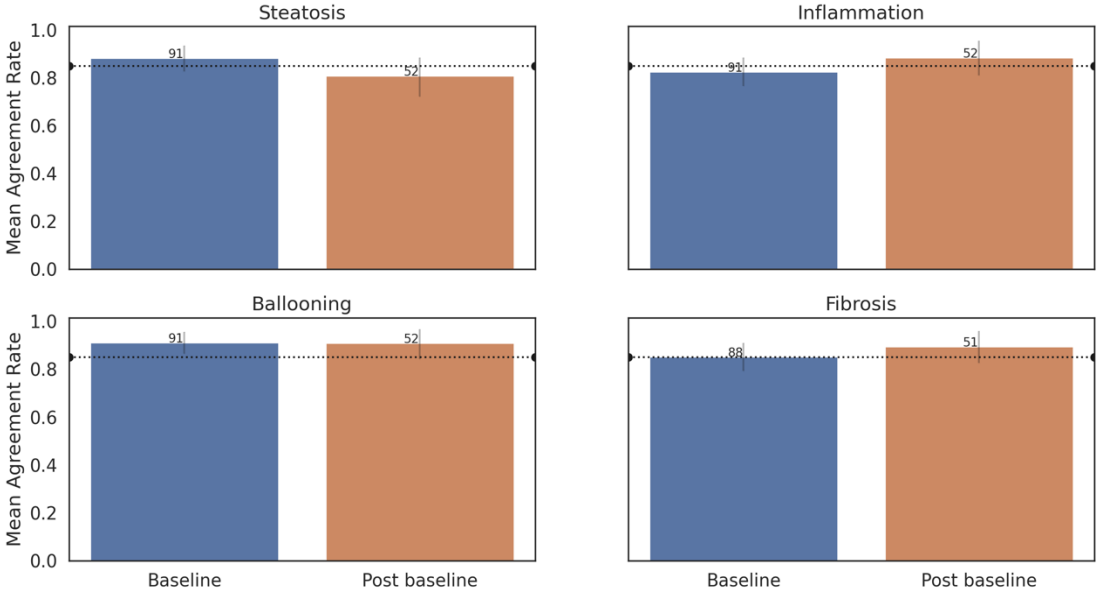


Table 62: Mean agreement rate by time point for reproducibility

Feature	Visit	N	Agreement, 95% CI
Steatosis	Baseline	91	0.883, (0.826, 0.934)
	Post baseline	52	0.808, (0.722, 0.884)
Lobular inflammation	Baseline	91	0.824, (0.765, 0.884)
	Post baseline	52	0.885, (0.81, 0.954)
Hepatocellular ballooning	Baseline	91	0.912, (0.865, 0.957)
	Post baseline	52	0.910, (0.843, 0.968)
Fibrosis	Baseline	88	0.852, (0.794, 0.91)
	Post baseline	51	0.895, (0.827, 0.96)

Predefined Exploratory Endpoints for Reproducibility

For exploratory endpoints, mean agreement rate for each score level of all NASH components was assessed between AIM-NASH scoring of inter-scanner/inter-operator WSIs (Figure 34 and Table 63). For steatosis, the mean agreement rate for scores demonstrated approximately 85% agreement, with mean agreement rates of 0.852 (95% CI of 0.759, 0.929), 0.848 (95% CI of 0.773, 0.911), 0.848 (95% CI of 0.777, 0.907), and 0.884 (95% CI of 0.815, 0.942) for scores 0, 1, 2, and 3, respectively. Alternatively, for fibrosis, there was a wider range of agreements, with fibrosis 3 mean agreement rate of 0.943 (95% CI, 0.895, 0.98) and fibrosis 4 mean agreement rate of 0.973 (95% CI, 0.942, 1). Notably, fibrosis 0 and 1, displayed the lowest agreement rates with more variability around the mean, but also represent a smaller sample size comparative to higher score levels. For ballooning, mean agreements were above 85%, with ballooning score 2 being significantly above 85% with a mean agreement of 0.949, (0.919, 0.976). For lobular inflammation score 2, the mean agreement was above 85% (0.886, (0.839, 0.928)), but other score levels were slightly below, noting that the sample size for inflammation 3 was only n=7.

Figure 34: Mean agreement by NASH score component for reproducibility

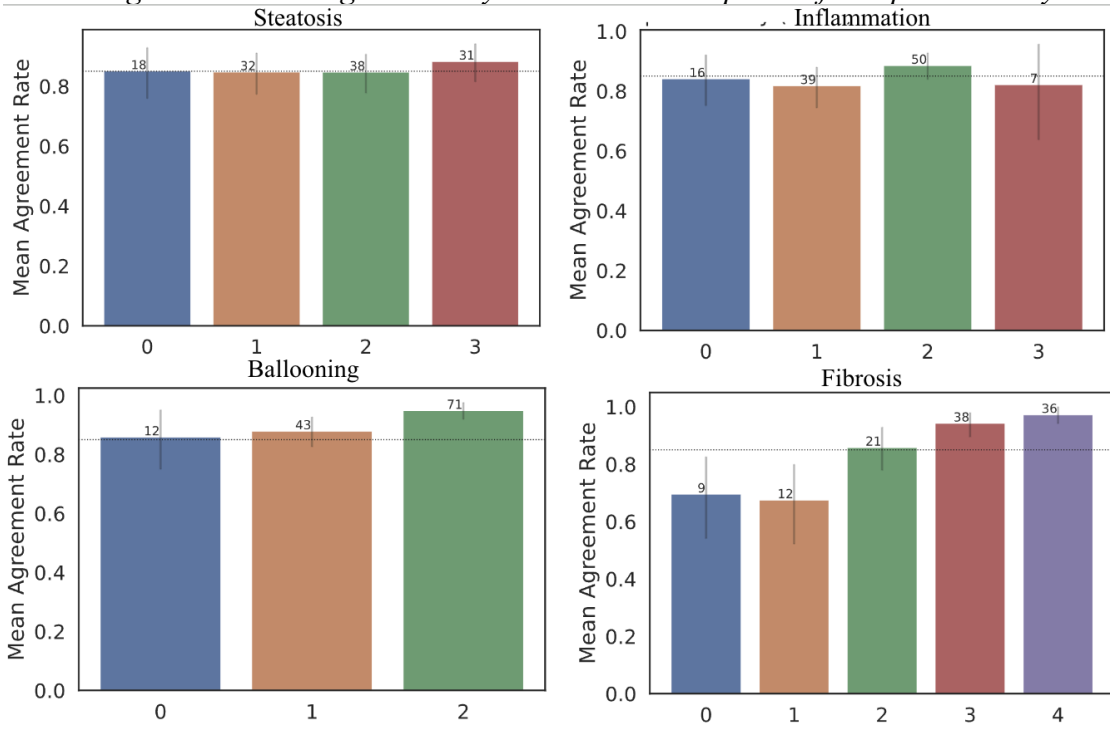


Table 63: Mean agreement rate by NASH score component for reproducibility

Measure	Score	N	Agreement, 95% CI
Steatosis	0	18	0.852, (0.759, 0.929)
	1	32	0.848, (0.773, 0.911)
	2	38	0.848, (0.777, 0.907)
	3	31	0.884, (0.815, 0.942)
Lobular inflammation	0	16	0.842, (0.751, 0.921)
	1	39	0.819, (0.744, 0.88)
	2	50	0.886, (0.839, 0.928)
	3	7	0.822, (0.637, 0.957)
Hepatocellular ballooning	0	12	0.859, (0.75, 0.951)
	1	43	0.880, (0.826, 0.927)
	2	71	0.949, (0.919, 0.976)
Fibrosis	0	9	0.696, (0.541, 0.826)
	1	12	0.676, (0.522, 0.799)
	2	21	0.859, (0.779, 0.929)
	3	38	0.943, (0.895, 0.98)
	4	36	0.973, (0.942, 1)

Exploratory analysis per trial of origin for each NASH component is shown in Figure 35 and Table 64. For hepatocellular ballooning, observed mean agreement rates for all trials were over 85%. For other components it was more variable, ranging from 0.765 to 0.926 for steatosis, 0.753 to 0.909 for lobular inflammation and 0.792 to 1.00 for fibrosis. All of these are still higher than published inter-reader agreements (Table 3).

Figure 35: Mean agreement by trial of origin for reproducibility

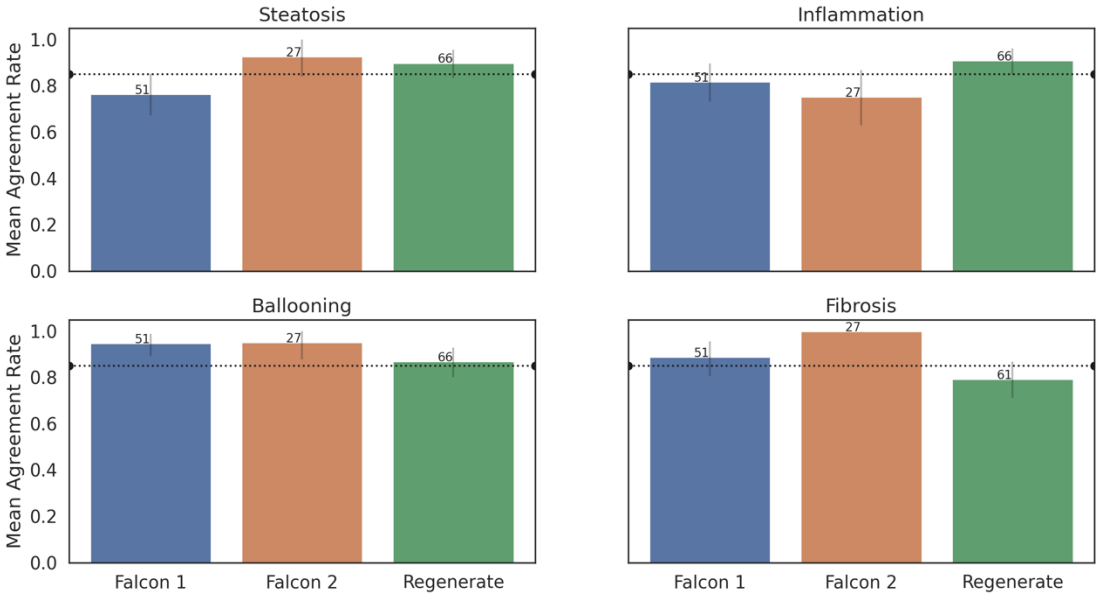


Table 64: Mean agreement rate by trial of origin for reproducibility

Feature	Trial	N	Agreement (95% CI)
Steatosis	Falcon 1	51	0.765 (0.674, 0.852)
	Falcon 2	27	0.926 (0.84, 1)
	Regenerate	66	0.899 (0.836, 0.956)
Lobular inflammation	Falcon 1	51	0.817 (0.733, 0.896)
	Falcon 2	27	0.753 (0.63, 0.867)
	Regenerate	66	0.909 (0.848, 0.961)
Hepatocellular ballooning	Falcon 1	51	0.948 (0.894, 0.988)
	Falcon 2	27	0.951 (0.878, 1)
	Regenerate	66	0.869 (0.802, 0.928)
Fibrosis	Falcon 1	51	0.889 (0.808, 0.955)
	Falcon 2	27	1.000 (0.875, 1)*
	Regenerate	61	0.792 (0.713, 0.867)

* Since the agreement rate is 1, the 95% CI is based on Wilson score confidence interval

4.5.11 Post hoc Exploratory Endpoints for Reproducibility

Post hoc exploratory analysis of mean agreement of AIM-NASH scoring on WSIs scanned on 3 different Aperio AT2 scanners in NASH aggregate scores (F0&F1 vs other, F4 vs other and NAS ≥4) were all significantly greater than 85% agreement (Figure 36,Table 65). These agreement rate for F0&F1 vs other was 0.957 (97% CI of 0.928, 0.981), F4 vs other was 0.986, (95% CI of 0.967, 1), and NAS ≥4 was 0.935 (95% CI of 0.9, 0.963).

Figure 36: Mean agreement for NAS aggregate scores for reproducibility

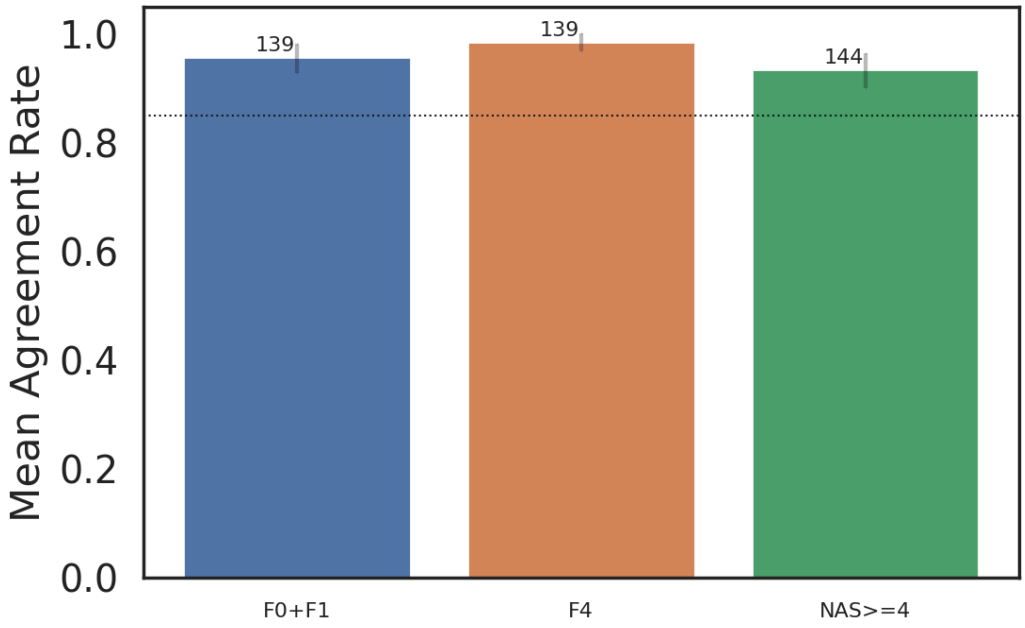


Table 65: Mean agreement rates for NASH aggregate scores for reproducibility

Feature	N	Agreement, 95% CI	p-value
F4 vs. other	139	0.986, (0.967, 1)	<0.0001
F0&F1 vs. other	139	0.957, (0.928, 0.981)	<0.0001
NAS ≥4 vs <4	144	0.935, (0.9, 0.963)	<0.0001

Finally, the pairwise inter-reader agreements were calculated between IMR pathologists across all cases (Table 66 and Table 67) in order to explicitly compare reproducibility across study pathologists to reproducibility achieved by AIM-NASH across sites and scanners. For all histologic components, AIM-NASH inter-scan, intra-site repeatability, and inter-scan, inter-site reproducibility was higher than for pathologist mean pairwise agreement (for pairs of pathologists who read at least 10 common cases) (Table 66).

Table 66: Manual pathologist vs. AIM-NASH repeatability and reproducibility

Feature	Mean AIM-NASH Inter-scan, Intra-site Repeatability (% Agreement	Mean AIM-NASH Inter-site Reproducibility (% Agreement)	Mean pairwise Agreement for Pathologists (% Agreement)
Steatosis	0.931	0.856	0.703
Lobular inflammation	0.958	0.847	0.453
Hepatocellular ballooning	0.963	0.912	0.556
Fibrosis	0.926	0.868	0.615

Table 67: Pairwise inter-reader agreement rates

Feature	Raters	N	Agreement Rate
Steatosis	Rater 1 vs Rater 2	606	0.632
	Rater 1 vs Rater 3	605	0.674
	Rater 1 vs Rater 4	80	0.725
	Rater 1 vs Rater 5	21	0.810
	Rater 1 vs Rater 6	10	1.000
	Rater 1 vs Rater 7	1	1.000
	Rater 2 vs Rater 3	605	0.590
	Rater 2 vs Rater 4	80	0.750
	Rater 2 vs Rater 5	21	0.762

Feature	Raters	N	Agreement Rate
	Rater 2 vs Rater 6	10	0.600
	Rater 2 vs Rater 7	1	1.000
	Rater 3 vs Rater 4	80	0.675
	Rater 3 vs Rater 5	21	0.714
	Rater 3 vs Rater 6	10	0.600
	Rater 3 vs Rater 7	1	1.000
	Rater 4 vs Rater 5	21	0.619
	Rater 4 vs Rater 6	10	0.600
	Rater 4 vs Rater 7	1	1.000
	Rater 5 vs Rater 6	10	0.800
	Rater 5 vs Rater 7	1	1.000
	Rater 6 vs Rater 7	1	1.000
Lobular inflammation	Rater 1 vs Rater 2	605	0.388
	Rater 1 vs Rater 3	605	0.397
	Rater 1 vs Rater 4	80	0.488
	Rater 1 vs Rater 5	21	0.429
	Rater 1 vs Rater 6	10	0.500
	Rater 1 vs Rater 7	1	0.000
	Rater 2 vs Rater 3	605	0.438
	Rater 2 vs Rater 4	80	0.238
	Rater 2 vs Rater 5	21	0.381
	Rater 2 vs Rater 6	10	0.800
	Rater 2 vs Rater 7	1	0.000
	Rater 3 vs Rater 4	80	0.388
	Rater 3 vs Rater 5	21	0.476

Feature	Raters	N	Agreement Rate
	Rater 3 vs Rater 6	10	0.600
	Rater 3 vs Rater 7	1	0.000
	Rater 4 vs Rater 5	21	0.476
	Rater 4 vs Rater 6	10	0.300
	Rater 4 vs Rater 7	1	1.000
	Rater 5 vs Rater 6	10	0.500
	Rater 5 vs Rater 7	1	1.000
	Rater 6 vs Rater 7	1	0.000
Hepatocellular ballooning	Rater 1 vs Rater 2	605	0.582
	Rater 1 vs Rater 3	605	0.469
	Rater 1 vs Rater 4	80	0.513
	Rater 1 vs Rater 5	21	0.429
	Rater 1 vs Rater 6	10	0.500
	Rater 1 vs Rater 7	1	0.000
	Rater 2 vs Rater 3	605	0.607
	Rater 2 vs Rater 4	80	0.450
	Rater 2 vs Rater 5	21	0.476
	Rater 2 vs Rater 6	10	0.800
	Rater 2 vs Rater 7	1	0.000
	Rater 3 vs Rater 4	80	0.563
	Rater 3 vs Rater 5	21	0.286
	Rater 3 vs Rater 6	10	0.600
	Rater 3 vs Rater 7	1	1.000
	Rater 4 vs Rater 5	21	0.667
	Rater 4 vs Rater 6	10	0.800

Feature	Raters	N	Agreement Rate
	Rater 4 vs Rater 7	1	0.000
	Rater 5 vs Rater 6	10	0.600
	Rater 5 vs Rater 7	1	0.000
	Rater 6 vs Rater 7	1	1.000
Fibrosis	Rater 1 vs Rater 2	604	0.606
	Rater 1 vs Rater 3	597	0.606
	Rater 1 vs Rater 4	77	0.519
	Rater 1 vs Rater 5	37	0.649
	Rater 1 vs Rater 6	16	0.750
	Rater 1 vs Rater 7	3	0.333
	Rater 2 vs Rater 3	597	0.637
	Rater 2 vs Rater 4	77	0.532
	Rater 2 vs Rater 5	37	0.649
	Rater 2 vs Rater 6	16	0.688
	Rater 2 vs Rater 7	3	0.333
	Rater 3 vs Rater 4	77	0.584
	Rater 3 vs Rater 5	37	0.514
	Rater 3 vs Rater 6	16	0.688
	Rater 3 vs Rater 7	3	0.000
	Rater 4 vs Rater 5	37	0.622
	Rater 4 vs Rater 6	16	0.688
	Rater 4 vs Rater 7	3	1.000
	Rater 5 vs Rater 6	16	0.500
	Rater 5 vs Rater 7	3	0.667
	Rater 6 vs Rater 7	3	0.333

4.5.12 Limitations

The slides used in this study were acquired from completed phase 2 and phase 3 clinical trials. The study was designed such that it is statistically powered with samples sizes able to observe differences by whole features, rather than by individual score levels and to provide evidence from multiple trials, demonstrating accuracy per the context of use population. The study is not powered at specific score levels within each histologic component, as this would have required significantly higher sample size and was not feasible to execute, and there are no published reference Kappas for manual pathology at each of these levels. However, exploratory analyses were performed to evaluate the success of the AIM-NASH at the individual score level compared to IMRs.

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 pathologists did not get to request a re-stain or a rescan of samples where 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 re-stained or rescanned.

4.5.13 Discussion and Conclusions

This AV study demonstrates the accuracy and precision of the AIM-NASH tool in measuring each component of the NAS score (steatosis, lobular inflammation, and hepatocellular ballooning) and CRN fibrosis stage on slides sourced from completed phase 2 and phase 3 NASH clinical trials. AIM-NASH accuracy in scoring steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis was compared to a GT panel of expert NASH pathologists and evaluated against independent manual pathologist readers (Figure 19). The AIM-NASH algorithm scores were not subject to pathologist review and represent the standalone algorithm assessment of NASH features. The sample set contained variability in disease state and severity, treatment with multiple drug candidates from both treatment and placebo arms, as well as variation in staining sourced from multiple central and local pathology labs used in the original trials. The scanning of glass slides and algorithm run execution for the study were performed by trained personnel in an external laboratory, and repeatability/reproducibility studies were performed across three laboratories. Meeting accuracy and reproducibility criteria in this diverse trial dataset would provide strong evidence that the algorithm is non-inferior for accuracy, highly reproducible, and superior to that demonstrated by manual pathology in the literature (6,14). This tool would thereby help to solve the current unmet need in using manual pathology, by ensuring accurate and consistent scoring during enrollment and more precise calculation of change over time for histologic-based endpoint analyses.

Accuracy analyses demonstrated that the AIM-NASH algorithm is superior to manual pathologist scoring for hepatocellular ballooning and non-inferior for steatosis, lobular inflammation, and fibrosis (Figure 21). Exploratory descriptive analyses demonstrated similar levels of accuracy on most score levels within histologic components with confidence intervals overlapping for algorithm and for manual reads, except for ballooning. Within ballooning, all scores were statistically superior in accuracy for AIM-NASH compared to manual scoring by individual pathologists (Figure 22). It is worth noting that AV samples sizes for some of the component scores were lower than others (e.g., lobular inflammation 0. 3 and fibrosis 0) due to availability of data in scores less represented in NASH trial populations. Regardless, accuracy was similar to the average expert pathologist performance at these score levels, and granularity at the score level illustrated in relevant literature datasets to our knowledge. As the screening and enrolled populations evolve with the NASH drug development field, there is potential for further improvement in NASH scoring, through additional training with improved methods and

expanded training populations. With the identification and scoring of hepatocellular ballooning being one of the most difficult aspects in CRN scoring and key in evaluating both disease activity and resolution of NASH, AIM-NASH results demonstrated here can greatly improve both scoring during enrollment and follow-up timepoints for endpoint evaluation.

In secondary and exploratory analyses, improved accuracy with AIM-NASH was also observed in specific NASH clinical trial relevant populations. Low stage fibrosis (F0&F1), fibrosis 2,3, and advanced fibrosis (F4) are useful benchmarks when considering NASH disease improvement or progression or during enrollment for cirrhotic and non-cirrhotic trials. Similarly, the sum of the ordinal scores for steatosis, inflammation, and ballooning being greater than or equal to 4 (NAS≥4) is one of the main indicators for a NASH diagnosis, as well as commonly being a requirement for trial inclusion. In addition, AIM-NASH demonstrated superiority (Table 54) over manual readers in evaluation of NASH resolution using the H&E slide (defined by ballooning=0, inflammation= 0 or 1, and any value for steatosis). In the current study AIM-NASH demonstrated significantly higher concordance with ground truth pathologists than independent manual readers for all of the evaluated trial inclusion and endpoint scenarios (Table 54 and Table 65). This supports implications for AIM-NASH as a powerful tool in NASH clinical trials, capable of accurately and consistently evaluating patients for trial inclusion criteria and in measuring the success or failure to meet study endpoints, which is currently an unmet need. Additional analysis was also performed to determine concordance of AIM-NASH with ground truth consensus pathologists at various timepoints throughout trials. Patient biopsies representing screening eligibility (including screen failures), baseline, and post-baseline follow-up were included in this study. Results showed equivalent or improved (hepatocellular ballooning) AIM-NASH concordance with ground truth pathologists, compared to independent manual readers (Table 48, Table 51, Table 52, Table 53, and Table 55). In line with the significance of the previous observation with clinical trial relevant NAS and fibrosis score groups, the ability of AIM-NASH to evaluate biopsies accurately and consistently during screen eligibility shows the potential of the tool to improve patient enrollment into prospective trials, as well as to accurately evaluate histologic-based endpoints.

Intra- and inter-reader variability when reviewing liver biopsies is a significant problem both for NASH trials and for the clinical care of biopsied patients. The potential to minimize these variables was tested through the repeatability and reproducibility arms of this AV. Percent agreement was calculated for algorithm repeatability (AIM-NASH run on the same WSI multiple times), scanner repeatability (AIM-NASH run on WSIs from glass slides scanned on multiple days within the same external site and same scanner operator), and scanner reproducibility (AIM-NASH run on WSIs from glass slides scanned at three external sites with different operators and scanners). The target of >0.85 per histologic component was set in order to exceed most intra- and inter-pathologist performance in reported literature. Percent agreement results for AIM-NASH repeatability was 100% for all NAS components and fibrosis when the algorithm was run five times on the same WSIs for the whole repeatability slide set (Table 60). For same site scanner repeatability (Table 50, percent agreement was significantly higher than 85% for each of the four NASH components. This high agreement persisted at the score level within each component, demonstrating the consistency of the algorithm more granularly across different scans from multiple days. At study timepoints of screen eligibility, baseline, and post-baseline follow-up, performance of the algorithm again surpassed the 0.85 threshold. These data suggest that the algorithm can successfully provide reliable NASH feature scores from WSIs produced on a given scanner across multiple days. This means, with well-defined scanner calibration-verification procedures (per College of American Pathologists

(CAP) guidelines), and image QC processes, the AIM-NASH can produce accurate scores, consistently throughout a NASH trial, which should increase precision in enrolling consistent populations and measuring change over time for trial endpoints.

For inter-site reproducibility, the glass slide set was scanned across three external sites (each a part of Covance) by different operators using different Aperio AT2 scanners. Overall, reproducibility was near 85% for all components, but lower bounds of the confidence intervals fell below 85% for steatosis, inflammation, and fibrosis (Figure 32 and Table 61). For hepatocellular ballooning, agreement was significantly above 85%. When describing reproducibility per timepoint, again agreement was approximately 85% but fell below for steatosis at post-baseline and inflammation at baseline (Figure 33 and Table 62). When looking at individual score agreements per component, again most values were around or above 85%, except for fibrosis which had larger variability in % agreement among the score levels, with a lower agreement for F0&1, but these scores also had substantially low sample numbers (Figure 36, Table 65). Similarly, agreement varied per trial sponsor, with steatosis being lower for Falcon 1, lobular inflammation being lower for Falcon 1 and 2, and fibrosis being lower for Regenerate (Figure 35, Table 64). It should be noted that the trichrome slides from Regenerate represented a wide variety of stain quality, many being faded and several years old. Reproducibility percent agreement was also calculated for some clinical trial-relevant subgroups (e.g., NAS $\geq$ 4), and AIM-NASH achieved significantly above 85% agreement for each (Figure 36, Table 65). Most importantly, overall, both scanner repeatability (inter-scan, intra-site) and reproducibility (inter-scan, inter-site) were higher than reported inter-pathologist agreements (14), and also substantially higher than mean pairwise agreement for pathologists' manual reads in this study (Table 61), thereby greatly improving consistency across reads and, of course, also reducing any temporal bias for future enrollment reads. For inter-site reproducibility, more detailed and standardized post-scanning image QC procedures (pre-algorithm run) could help to achieve even higher levels of reproducibility. In most cases, however, trial slides are scanned at one site, where the scanner is maintained via CAP/CLIA guidelines throughout all trial work, and the algorithm has been demonstrated here to be highly repeatable (>90%) across scans from multiple days in this diverse AV dataset.
Overall, these data present the AIM-NASH algorithm alone as an accurate, robust tool with strong potential for driving more standardized, rigorous, and consistent clinical trial enrollment, monitoring, and therefore more accurate determination of histologic change over time for trial endpoints.

4.6 Overlay Validation

4.6.1 Study Purpose

The purpose of Overlay Validation is to evaluate the accuracy of the AI-generated overlays in terms of true positives and false positives.

4.6.2 Objectives

Background: In addition to the machine learning-derived scores, the AIM-NASH user interface also displays the WSI, with overlays, corresponding to the scores the tool has predicted. The pathologist has the option of displaying the colored overlays showing the CNN model's predictions of areas containing the histologic features of interest (i.e., steatosis, lobular inflammation, and ballooning for H&E-stained tissue, fibrosis for trichrome-stained tissue). In initial user studies, pathologists reported utility in viewing the overlays to "highlight" these features, and would then assess them at higher magnification, toggling on/off the overlay to assess the

morphology, which could potentially result in creating efficiencies in evaluating key histological features when scoring and reviewing the AIM-NASH score. Therefore, the overlays' intended use is as an assist tool, guiding the pathologist to the locations of relevant histologic features in each slide.

The objective of the overlay validation is to assess the accuracy of machine learning-derived heatmap overlays in highlighting steatosis, lobular inflammation, ballooning, and fibrosis, as well as slide and scanning artifacts, in WSIs of NASH biopsies stained with H&E or trichrome.

Up to one hundred and sixty (160) 500 x 500-micron-sized frames (sampled from NASH biopsy WSIs) will be enrolled for each feature being evaluated. Frames will be enrolled to evaluate one feature or multiple features. These enrolled frames will display the AIM-NASH model's predictions for steatosis, lobular inflammation, ballooning, and detected glass slide and digital artifacts over the H&E-stained image, and fibrosis and detected glass slide and digital artifacts, over the trichrome-stained image, depending on which feature the frame is selected to be evaluated for.

Study Objective: The primary objective of this study was to assess the accuracy of machine learning-derived overlays in highlighting steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis, as well as slide and scanning artifacts, in WSIs of NASH biopsies stained with H&E or trichrome.

4.6.3 Study Design and Plan

For this study, up to 160 500 x 500-micron-sized frames were to be enrolled for each feature being evaluated. These frames were sampled from NASH biopsy WSIs from slides from the same clinical trial data sets that were utilized for the analytical and clinical validation studies (Intercept Regenerate trial, Bristol Myers Squibb Falcon 1 and Falcon 2 trials and Novo Nordisk Semaglutide Phase 2, non-cirrhotic NASH trial).

For the enrolled frames to hold a representative distribution of the features, the AIM-NASH heatmaps were used as an approximate guide for algorithmically identifying frames with relevant tissue. During frame sampling an approximate 2:1 ratio was followed in which twice as many frames were sampled than were enrolled in the trial. Using this ratio for frame sampling ensures the presence of appropriate features and mitigates sampling bias by making certain frames enrolled into the study are done so using independent pathologist metrics and not the AIM-NASH heatmaps under validation. The AIM-NASH algorithm was utilized as a pre-screen to select the frames for the Frame Enrollment Task.

For the Frame Enrollment Task, a qualified PathAI Contributor Network pathologist who has demonstrated extensive experience reviewing NASH biopsies and was not involved in AIM-NASH model development will perform frame selection. For each frame, the pathologist will be asked to identify how much of a particular feature is present. Specifically, the enrolling pathologist reviewed the following questions for each feature:

- **Steatosis**
  - Task 1 – What percentage of the frame area is covered by steatosis [0-100%]?
- **Artifact**
  - Task 1 – What percentage of the frame area is covered by artifact [0-100%]?
- **Fibrosis**
  - Task 1 – What percentage of the frame area is covered by fibrosis [0-100%]?
  - Task 2 – Does this frame have little or no liver parenchyma (large portal tracts, all capsule/septum)?

        - Yes
        - No
    - **Ballooning**
        - Task 1 – How many cells of ballooning are present within the frame?
            - None
            - 1-Few
            - Frequent
    - **Lobular Inflammation**
        - Task 1 – Roughly how many foci of lobular inflammation are present within the frame?
            - None
            - 1
            - 2-4
            - >4

These enrollment evaluations were made without the use of the AIM-NASH overlays. Based on this input, frames were enrolled into specific feature buckets by an unblinded PathAI clinical data manager to fulfill the target distribution ranges in Table 68.

Table 68: Approximate Distribution Requirements of the Frame Evaluation Task

% Steatosis in a Frame (# of frames)	# of Inflammation Foci in a Frame (# of frames)	# of Ballooning Cells in a Frame (# of frames)	% Fibrosis in a Frame (# of frames)	% H&E Artifact in a Frame (# of frames)	% Trichrome Artifact in a Frame (# of frames)
None (8–16)	None (8–16)	None (8–16)	None (8–16)	None (8-32)	None (8-32)
Low (24–84)	1 (24–84)	1–Few (32–120)	Low (24–84)	Artifact (128-152)	Artifact (128-152)
Medium (24–84)	2–4 (24–84)	Frequent (32–120)	Medium (24–84)	X	X
High (24–84)	>4 (24–84)	X	High (24–84)	X	X

Frames were enrolled by the pathologist by reviewing 240 frames at a time until each desired distribution category was filled. Five enrollment rounds were required to meet the distributions.

Each frame also has ground truth (GT) scores to ensure that the frames came from slides with a variety of scores. The source of ground truth is the consensus pathologist score collected by PathAI Contributor Network pathologists. The consensus reads are performed by two panels of two pathologists, with a third tiebreaker pathologist, where necessary. The pathologists are chosen based on their previous experience and proficiency results of previously completed slides for PathAI (see Appendix IIc for more detail). For GT, when the two primary reads agreed, that score was designated as the overall GT score for that particular component. In cases where the two primary reads disagreed on any of the NASH component scores (steatosis, lobular inflammation, hepatocellular ballooning and/or fibrosis), the slide was sent out to a third, tiebreaker pathologist. To blind the tiebreaker to which component was discordant, all NAS components were sent out for scoring. If the tiebreaker pathologist agreed with one of the primary pathologists on the discrepant score(s), this was then determined to

be the final score for that NASH component. If the tiebreaker pathologist disagreed with both primary pathologists, a joint panel call was held with the three pathologists to come to a consensus on the discrepant component(s), with the tiebreaker providing the final score in the rare case that consensus could not be reached. The tiebreaker pathologist was the same for both panels. Overall, 5 pathologists provided scores for the GT reads.

Once the frames were enrolled, they were sent out to 3 board-certified expert hepatopathologists, who were trained on the study protocol. The 500 x 500-micron-sized frames (sampled from NASH biopsy WSIs) displayed the model's overlay predictions for steatosis, lobular inflammation, ballooning, as well as artifacts over the H&E-stained image, and fibrosis and detected glass slide and digital artifacts, over the trichrome image (where, additionally, "artifacts" that were masked by the algorithm were defined). The 3 evaluation pathologists were asked specific questions (See Appendix VIIc for Example Case Report Forms) for each frame to determine if the overlay was capturing a critical proportion of the true feature of interest in the frame (true positives) and whether the model was predicting a particular feature where the feature was not present (false predictions).
In parallel, and as part of the AIM-NASH CV studies, expert pathologists are asked to qualitatively comment on the utility of the heatmap overlays in the context of accepting or rejecting the AIM-NASH scores for each of the four key histologic features. See Appendix IVa for the Clinical Validation Protocol.

4.6.4 Dataset

Whole slide images were selected from those also available for use in AV and CV for the AIM-NASH DDT overlay validation. The slides were selected from four completed NASH phase 2b or 3 clinical trials (Table 69). Slides were selected such that they reflect a representative distribution of disease severity.

Table 69: Clinical Trials Used for Slide Selection

Trial Name and Sponsor	Trial Phase	Drug	Enrollment Criteria
REGENERATE Intercept Pharmaceuticals	3	Obeticholic Acid	Presence of all 3 NAS components Fibrosis stage 2 or stage 3 OR Fibrosis stage 1a or stage 1b if accompanied by ≥1 of the following risk factors: Obesity (BMI ≥30 kg/m2) Type 2 diabetes diagnosed per 2013 American Diabetes Association criteria ALT >1.5× upper limit of normal (ULN).
FALCON2 Bristol Myers Squibb	2	Pegbelfermin	Biopsy must be consistent with NASH Biopsy must be consistent with cirrhosis (stage 4)
FALCON 1 Bristol Myers Squibb	2	Pegbelfermin	A score of at least 1 for each NASH component Fibrosis Stage 3
Semaglutide NASH Trial (NCT04822181) Novo Nordisk	2	Semaglutide	Biopsy-proven NASH; A histological NAFLD activity score equal to or above 4 with a score of 1 or more in steatosis, lobular inflammation and hepatocyte ballooning Fibrosis stage 2,3

4.6.5 Selection of Study Population/ Cases

Inclusion Criteria

- De-identified NASH biopsy WSIs from phase 2 and phase 3 NASH clinical trials.
- Samples where consent for use for additional research was collected at the time of initial enrollment.
- Only frames from H&E and trichrome blue WSI.
- Frames enrolled based on the predetermined distribution shown in Table 68

Exclusion Criteria

- WSIs from any other trichrome stain (e.g., trichrome green).
- Cases used for model development.
- Cases that have any patient identifying information.
- Cases with other than NASH FFPE tissue.
- Cases with little or no liver parenchyma.

4.6.6 General Procedures

Blinding: All participating pathologists had their own unique log in to the PathAI digital slide viewer and they were assigned jobs with relevant frames. The pathologists were blinded to each other's assessments and the enrollment pathologist was also blinded to the AIM-NASH overlays. All PathAI staff (except for the unblinded clinical data managers and unblinded clinical scientist) involved in this study were blinded to the data until the database was locked.

4.6.7 Pathologist Training

All pathologists were trained on study protocol and required tasks by the principal investigator (PI) prior to participating in any study activities. All pathologists also signed an attestation form acknowledging the completion of training. All training records are stored in PathAI's electronic Quality Management System (eQMS). In addition, prior to the start of the study, each pathologist completed 30 training frames on PathAI's digital viewer platform to familiarize themselves with and confirm understanding of the study task(s).

4.6.8 Data Handling

All data was entered electronically in the PathAI digital viewer platform and after the completion of the study, all data was downloaded from the platform and stored in PathAI's eQMS. PathAI designated clinical data managers and a clinical scientist who were unblinded to the data and had access to all study information for the purpose of monitoring data and resolving any queries.

PathAI designated clinical data managers and a clinical scientist were unblinded to the data and had access to all study information for the purpose of monitoring data and resolving queries.
All data was collected electronically in the PathAI digital viewer platform. Data was downloaded from the database service for the PathAI digital viewer platform over the course of the study for data monitoring purposes. All relevant study data along with corresponding documentation was uploaded to the Clinical Data Management AWS bucket which only unblinded clinical data managers have access to.

4.6.9 Statistical Methods and Determination of Sample Size

Primary Analysis: Overlay performance was considered acceptable if true positive (TP) success rate and false positive (FP) success rate were greater than or equal to 85%.

- **True Positive (TP; measure of sensitivity) success (Question 1):**
  - Macrovesicular Steatosis + Fibrosis
    - The feature is present in the frame and less than or equal to 10% of the total frame area is being underestimated by overlay.
  - Ballooning + Lobular Inflammation
    - The feature is present in the frame and overlay is sufficiently identifying feature to report a correct grade for the frame.
  - Artifact
    - Artifact is present in the frame and overlay is identifying artifact
- **False Positive (FP; measure of specificity) success (Question 2):**
  - Less than or equal to 20% of the total frame area is being overestimated by overlay.

For each success criteria, an overall success rate was calculated as the mean of fraction of frames deemed to have met the success criteria aggregated across the three evaluating pathologists, i.e. [((P1_N_success/P1_N_total) + (P2_N_success/P2_N_total) + (P3_N_success/P3_N_total)) / 3] where Px_N_success is the number of frames meeting the TP success or FP success, as described above, for each pathologist. Additionally, Px_N_total is the number of frames deemed as containing each substance for evaluating TP and FP success. This number can be different across the pathologists as they can differ in their assessment of each substance.
Separately for each feature under review, TP success rate and FN success rate was assessed as aggregated across the three rating pathologists. Performance for each NASH feature and artifact was accepted if the lower 2.5% CI for TP success rate and FN success rate was above 85%.
Scores for each rating pathologist for each feature, and their 95% CI, is also presented.

Bootstrap resampling with replacement was used to compute confidence intervals across K=2000 full sample-size replicates. 95% confidence interval was defined as the 2.5^{th} and 97.5^{th} percentiles of the bootstrap distribution. In cases where the success rate is 100%, confidence interval will be computed using Wilson score method instead of bootstrap.

Exploratory Analysis: The exploratory analysis examined sources of variability between pathologists for features with more discordance between pathologists in identifying slides with that feature.

Determination of Sample Size: Assuming a success criterion of 95% and a target of exceeding 85% at one-sided alpha=0.025 level. Wilson score confidence intervals then gives N (for each overlay) of 115 or 138 at 90% or 95% power, respectively. To ensure adequate power, and account for the possibility of missing features, the N for each feature was chosen at least 160 frames.

4.6.10 OV Results

Final Study Population

Frames from 222 digital slides were enrolled. The slides were from completed clinical trials (Intercept Regenerate trial, Bristol Myers Squibb Falcon 1 and Falcon 2 trials and Novo Nordisk Semaglutide trial). Five enrollment rounds were performed. For 4 rounds, the 240 frames were split up between H&E and trichrome slides. For the fifth enrollment round, 38 frames from trichrome WSIs were utilized. One hundred and sixty frames were enrolled

for each feature, and some frames were utilized for multiple features. Overall, 312 unique H&E frames and 249 trichrome frames were enrolled.

Data Sets Analyzed

The datasets included in the overlay validation study were the same as for AIM-NASH analytical and clinical validation studies and included two phase 2 studies from Bristol Myers Squibb (BMS; Falcon 1 and Falcon 2), one phase 2 study from Novo Nordisk (NN9931-4296 Semaglutide trial) and one phase 3 study from Intercept (Regenerate).

Demographic and Other Baseline Data

No demographic information for the slides enrolled in the study is available. The dataset contains a broad spectrum of disease presentation, represents both screened and enrolled patient populations, including study subjects who may have regressed or progressed during a clinical trial, and reflects the NASH clinical trial population. The dataset also contains variability in sample collection (historical biopsies vs. study biopsies), staining (including performed by multiple collection/preparation sites), and populations representing various treatments with candidate therapies. Distribution of frames based on slide level score (ground truth scores) are listed in Table 70 and distribution of frames based on frame level scores (collected from the enrollment pathologist) are listed in Table 71. Slide and frames level distributions per sponsor (BMS, Novo Nordisk and Intercept) are listed in
Table 72 and Table 73.

Table 70: Frame Distribution based on Slide Level Score

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

Note: N<160 since multiple frames can selected from a slide

Table 71: Frame Distribution based on Frames Level Score

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

Table 72: Distribution of Slides Based on Sponsor

Feature	Sponsor	(n/N)	%
H&E Artifact	BMS	(28/72)	38.9
	Intercept	(18/72)	25.0
	Novo Nordisk	(26/72)	36.1
Hepatocellular ballooning	BMS	(28/86)	32.6
	Intercept	(22/86)	25.6
	Novo Nordisk	(36/86)	41.9
Lobular inflammation	BMS	(30/87)	34.5
	Intercept	(23/87)	26.4
	Novo Nordisk	(34/87)	39.1
Steatosis	BMS	(30/87)	34.5
	Intercept	(22/87)	25.3
	Novo Nordisk	(35/87)	40.2
Trichrome Artifact	BMS	(14/62)	22.6
	Intercept	(15/62)	24.2
	Novo Nordisk	(33/62)	53.2
Fibrosis	BMS	(19/79)	24.1
	Intercept	(28/79)	35.4
	Novo Nordisk	(32/79)	40.5

Note: N<160 since multiple frames can selected from a slide

Table 73: Distribution of Frames Based on Sponsor

Feature	Sponsor	(n/N)	%
H&E Artifact	BMS	(57/160)	35.6
	Intercept	(44/160)	27.5
	Novo Nordisk	(59/160)	36.9
Hepatocellular ballooning	BMS	(59/160)	36.9
	Intercept	(40/160)	25.0
	Novo Nordisk	(61/160)	38.1
Lobular inflammation	BMS	(60/160)	37.5
	Intercept	(40/160)	25.0
	Novo Nordisk	(60/160)	37.5
Steatosis	BMS	(60/160)	37.5
	Intercept	(40/160)	25.0
	Novo Nordisk	(60/160)	37.5
Trichrome Artifact	BMS	(63/160)	39.4
	Intercept	(44/160)	27.5
	Novo Nordisk	(53/160)	33.1
Fibrosis	BMS	(54/160)	33.8
	Intercept	(39/160)	24.4
	Novo Nordisk	(67/160)	41.9

Primary Analysis

A total of 160 frames per feature (steatosis, lobular inflammation, hepatocellular ballooning, fibrosis, H&E artifact, and trichrome artifact) were evaluated in this study. Frames were enrolled in the study by one enrolling pathologist without the use of AIM-NASH overlays. The enrolling pathologist was not involved in the evaluation task. AIM-NASH overlays for each enrolled frame were evaluated by 3 qualified hepatopathologists. For each frame and each feature, the pathologists indicated whether the feature was present (yes/no). The presence of each feature per pathologist is shown in Table 74. The highest degrees of variability in the presence/absence of a feature in a frame were observed for ballooning (feature present in 57.5%, 44.4% and 69.4% of the frames per pathologist A, B and C, respectively), inflammation (feature present in 82.5%, 82.5% and 96.9% of the frames per pathologist A, B and C, respectively) and for trichrome artifact (feature present in 71.3%, 77.5% and 93.1% of the frames per pathologist A, B and C, respectively).

Table 74: Presence of Feature per Pathologist

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

The acceptance criteria for TP (underestimation) success were met for all feature overlays except for ballooning (Figure 37 and Table 75). H&E artifact TP success rate was 0.97 (95% CI, 0.95, 0.992), trichrome artifact was 0.99 (95% CI, 0.968, 1), inflammation was 0.94 (95% CI, 0.915, 0.962), steatosis 0.96 (95% CI, 0.932, 0.98) and fibrosis 0.97 (95% CI, 0.946, 0.988). For ballooning the overall TP success rate was 0.87, with 95% CI (0.8333,0.913).

Figure 37: Overall True Positive Success Rate

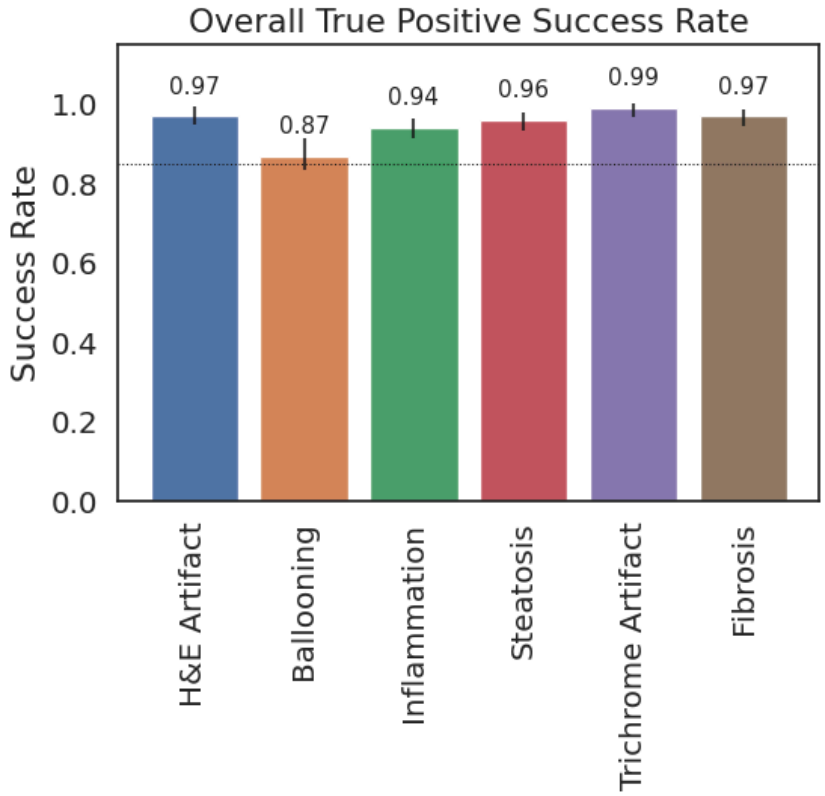


Table 75: True Positive Success Rates per Overlay Feature

Feature	Success Rate	95% CI
H&E Artifact	0.97	(0.95, 0.992)
Hepatocellular ballooning	0.87	(0.833, 0.913)
Lobular inflammation	0.94	(0.915, 0.962)
Steatosis	0.96	(0.932, 0.98)
Trichrome Artifact	0.99	(0.968, 1)
Fibrosis	0.97	(0.946, 0.988)

The individual pathologist TP success rates for each overlay feature are also shown in Figure 38 and listed in Table 76. The individual pathologist TP success rates show variability for ballooning overlay, where pathologist A and B TP success rates are 0.96 (95% CI, 0.911, 0.991) and 0.94 (95% CI, 0.89, 0988) respectively. However, pathologist C TP success rate for ballooning overlay was only 0.72 (95% CI, 0.639, 0.805). Pathologist C TP success rate for inflammation overlay was also lower than for pathologists A and B with TP success rates of 0.86 (95% CI, 0.809, 0.919), 0.98 (95% CI, 0.95, 1) and 0.98 (95% CI, 0.947, 1), respectively.

Figure 38: Individual Pathologist True Positive Success Rate

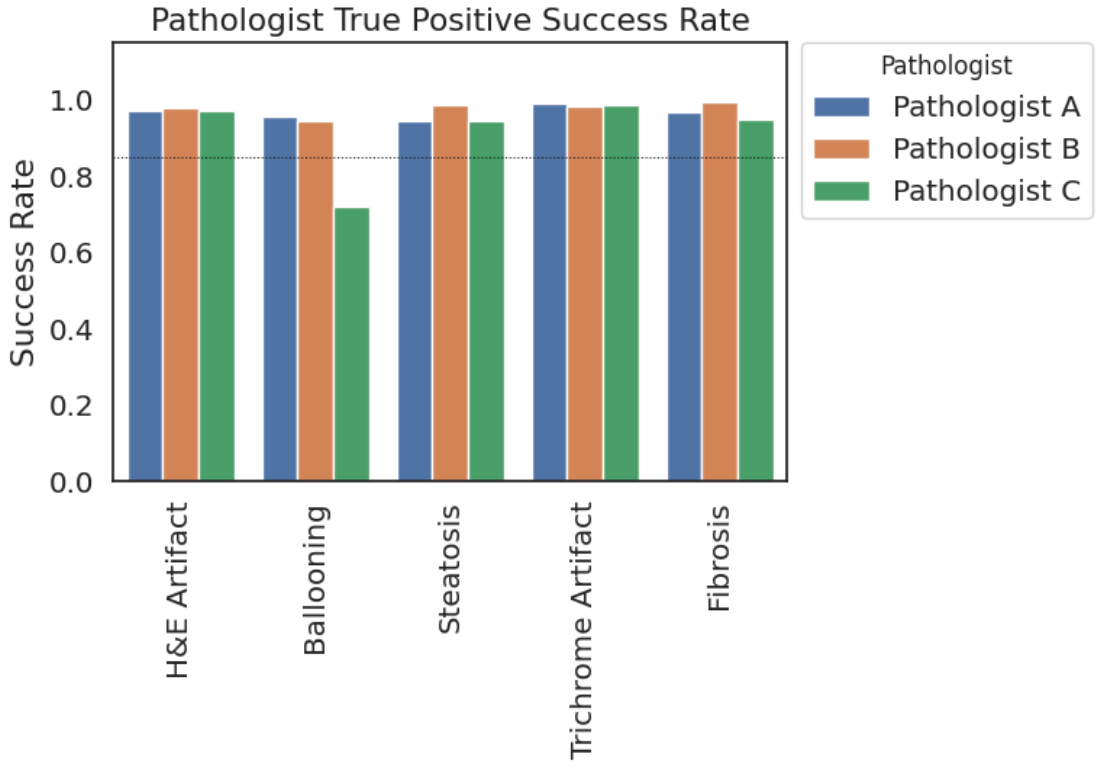


Table 76: True Positive Success Rate per Individual Pathologist

Feature	Pathologist	Success Rate	95% CI
H&E Artifact	A	0.97	(0.943, 0.993)
	B	0.98	(0.95, 1)
	C	0.97	(0.942, 0.993)
Hepatocellular ballooning	A	0.96	(0.911, 0.991)
	B	0.94	(0.89, 0.988)
	C	0.72	(0.639, 0.805)
Lobular inflammation	A	0.98	(0.95, 1)
	B	0.98	(0.947, 1)
	C	0.86	(0.809, 0.919)
Steatosis	A	0.94	(0.904, 0.976)
	B	0.99	(0.967, 1)
	C	0.94	(0.904, 0.976)
Trichrome Artifact	A	0.99	(0.972, 1)
	B	0.98	(0.958, 1)
	C	0.99	(0.967, 1)
Fibrosis	A	0.97	(0.935, 0.993)
	B	0.99	(0.979, 1)
	C	0.95	(0.909, 0.98)

The acceptance criteria for FP (overestimation) success rate were met for all 6 feature overlays (Figure 39 and Table 77). H&E artifact success rate for FP was 0.973 (95% CI, 0.948, 0.992), trichrome artifact was 0.931 (95% CI, 0.9, 0.958), inflammation was 0.992 (95% CI, 0.983, 0.998), steatosis was 1.00 (95% CI, 0.977, 1), ballooning was 0.921 (95% CI, 0.899, 0.942) and fibrosis was 0.998 (95% CI, 0.993, 1).

Figure 39: Overall False Positive Success Rate

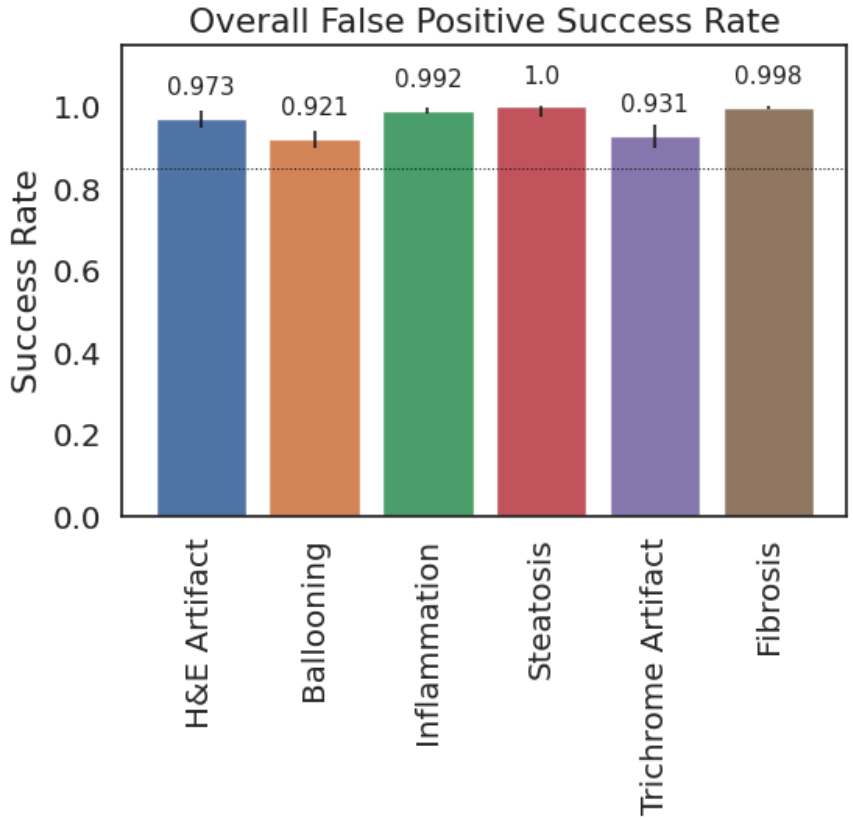


Table 77: False Positive Success Rates per Overlay Feature

Feature	Success Rate* (True Negative)	95% CI
H&E Artifact	0.973	(0.948, 0.992)
Hepatocellular ballooning	0.921	(0.899, 0.942)
Lobular inflammation	0.992	(0.983, 0.998)
Steatosis	1.00	(0.977, 1)
Trichrome Artifact	0.931	(0.9, 0.958)
Fibrosis	0.998	(0.993, 1)

*False positive rate is (1-False Positive Success Rate)

The individual pathologist FP success rates for each overlay feature are also shown in Figure 40 and listed in Table 78. The individual pathologist FP success rates again show the greatest variability for ballooning overlay, where pathologist A and B FP success rates are 1.00 (95% CI, 0.977, 1) and 1.00 (95% CI, 0.977, 1) respectively. However, pathologist C FP success rate for ballooning overlay was only 0.76 (95% CI, 0.696, 0.827).

Additionally, trichrome artifact overlay showed some variability in individual pathologists' FP success rates, where pathologist B and C FP success rates are 0.94 (95% CI, 0.906, 0.976) and 0.98 (95% CI, 0.948, 994) respectively. Pathologist A FP success rate for trichrome artifact overlay was slightly lower 0.88 (95% CI, 0.822, 0.924).

Figure 40: Individual Pathologist False Positive Success Rate

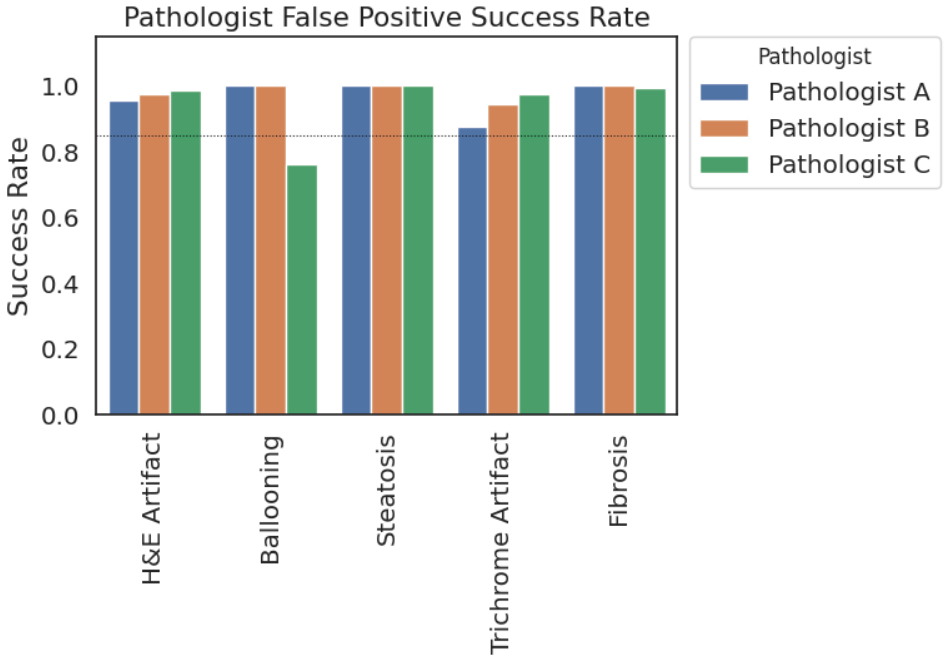


Table 78: False Positive Success Rate per Individual Pathologist

Feature	Pathologist	Success Rate	95% CI
H&E Artifact	A	0.96	(0.921, 0.987)
	B	0.98	(0.947, 0.994)
	C	0.99	(0.966, 1)
Hepatocellular ballooning	A	1.00	(0.977, 1)
	B	1.00	(0.977, 1)
	C	0.76	(0.696, 0.827)
Lobular inflammation	A	1.00	(0.977, 1)
	B	1.00	(0.977, 1)
	C	0.98	(0.948, 0.994)
Steatosis	A	1.00	(0.977, 1)
	B	1.00	(0.977, 1)
	C	1.00	(0.977, 1)
Trichrome Artifact	A	0.88	(0.822, 0.924)
	B	0.94	(0.906, 0.976)
	C	0.98	(0.948, 0.994)
Fibrosis	A	1.00	(0.977, 1)
	B	1.00	(0.977, 1)
	C	0.99	(0.979, 1)

Exploratory Analysis Results

For exploratory analysis variability between pathologists' identification of each feature was determined. Table 79 shows the proportion of frames where all 3 evaluating pathologists agreed on presence of the feature when at least 1 pathologist indicated presence of feature in a frame. The agreement for presence of hepatocellular ballooning was the lowest (55.1%) out of all features and therefore, sources of variability between pathologists for hepatocellular ballooning were further examined.

Table 79: Three Pathologist Agreement on Presence of Features in a Frame Where At Least One Pathologist Indicated Presence

Feature	(n/N) *	%
H&E Artifact	(132/148)	89.2
Hepatocellular ballooning	(65/118)	55.1
Lobular inflammation	(124/155)	80.0
Steatosis	(158/159)	99.4
Trichrome Artifact	(108/150)	72.0
Fibrosis	(149/154)	96.8

*n= number of frames where all 3 pathologists agreed on presence of the feature. N= number of frames where at least 1 pathologist indicated presence of the feature

For the 65 frames where all 3 evaluating pathologists indicated presence of hepatocellular ballooning, the TP success rate was calculated (Table 80). Pathologists A and B identified underestimation in 1 and 3 of the 65 frames, respectively, making their TP success rates 0.99 for pathologist A and 0.95 for pathologist B. However, pathologist C identified underestimation in 10 of the 65 frames, showing a TP success rate of 0.85.

Table 80: True Positive Success Rate for Hepatocellular Ballooning for Frames Where All 3 Pathologists Indicated Presence of Ballooning

Pathologist	Success Rate	(n/N)
A	0.985	(64/65)
B	0.954	(62/65)
C	0.846	(55/65)

4.6.11 Limitations

The slides utilized in this study come from completed phase 2 and 3 clinical trials and therefore have a limited number of slides with certain scores (e.g., fibrosis 0 is not prevalent in enrolled study populations). Because overlay validation is at frames level and not at slide level, the issue was mitigated by utilizing several frames from the limited number of slides available to meet the target distribution ranges.

In addition, some of the less prevalent features (artifact, lobular inflammation, and hepatocellular ballooning) only occupy up to 20% of the frame (and frequently less), making it difficult to reliably estimate of % frame area occupied by the feature. The overlay evaluation questions were designed with this in mind to obtain meaningful estimates from the pathologists. The evaluation overlay questions were also designed to follow the NASH scoring

guidelines for H&E slides, where steatosis is defined by percentage of feature present, but lobular inflammation and hepatocellular ballooning are more qualitative features and defined not be % area but as number of foci and none, few, or many, respectively.

4.6.12 Discussion and Conclusion

This AIM-NASH overlay validation study demonstrates that the tissue model overlays of AIM-NASH are accurate in highlighting steatosis, lobular inflammation, hepatocellular ballooning, fibrosis, H&E artifact, and trichrome artifact. These overlays are 3-dimensional matrices with dimensions corresponding to X-coordinate, Y-coordinate, and model outputs for each pixel in the digital images. The overlays are meant to serve as a highlighting tool to assist pathologists in identifying regions containing specific features, thereby creating efficiencies in evaluating key histological features when scoring and reviewing NASH scores. Additionally, the overlays can be toggled off and on to facilitate review based on pathologist preference.

The results of this study show that the overlays are accurate as a spotlight to highlight the features on NASH slides. Overlays for all features for TP and FP success rate met their acceptance criteria except for TP rate for ballooning where the 95 % CI was (0.833, 0.913). These results for lower TP for hepatocellular ballooning are not surprising, especially since there are widely varying approaches and interpretations, even among experienced hepatopathologists, regarding which specific cells constitute as ballooned hepatocytes (26). As a good example, the 2022 Brunt paper (26) demonstrates how only one cell among a large panel of candidates, was unanimously identified as a ballooned hepatocyte by 9 hepatopathologists who were annotating the same slide. Similar interobserver variation can be seen in our results when looking at presence of ballooning in a frame per pathologist, where pathologist B identified 44.4% of frames analyzed for ballooning having the feature present, pathologist A identified 57.5% of these frames as having ballooning and pathologist C as 69.4% having ballooning. This is a difference of 40 frames out of 160 where pathologist C identified ballooned hepatocytes and pathologist B did not. Additionally, from the exploratory analysis performed, it is likely that pathologist C identifies a sub-type of cells as ballooned hepatocytes that the other 2 do not. The difference in identifying ballooning cells by pathologist C and identifying underestimation of ballooning by AIM-NASH ballooning overlay by the same pathologist could be related and contributing to the slightly lower TP success rate for the ballooning overlay. This highlights the importance of training and harmonization, where possible, among trial pathologists, before a trial begins and as needed, during the progress of the trial. AIM-NASH workflow helps to standardize the identification of these complex features in a way which allows for accurate and consistent scoring, facilitated in part by the overlays, which will be further demonstrated by AV and CV studies.

Additionally, steatosis, inflammation, ballooning, and fibrosis overlays are also indirectly validated as part of the AV and CV studies of the AIM-NASH algorithm. These feature overlays serve as part of the input for GNNs that generate slide level scores, which are validated in AV and CV for accuracy, where CV additionally incorporates the pathologist review of the slide, the overlays, and the corresponding algorithm scores. As mentioned above, the overlays are not a main decision tool for the AIM-NASH workflow pathologist when reviewing the algorithm scores but rather a highlighting tool in order to increase efficiencies of AIM-NASH score review.

Overall, we conclude that the AIM-NASH overlay features are accurate in highlighting steatosis, lobular inflammation, hepatocellular ballooning, fibrosis, H&E artifact, and trichrome artifact as demonstrated by this overlay frames validation study.

4.7 Clinical Validation

4.7.1 Study Purpose

The purpose of clinical validation is to measure the ability of AIM-NASH to assist pathologists in their assessment of NASH as the tool would be utilized in a therapeutic trial setting.

4.7.2 Objectives and Endpoints

Objective	Endpoint
Primary To evaluate performance of AIM-NASH assisted pathologists when the tool is utilized in a NASH clinical trial setting.	**Primary** Accuracy of the AI-assisted workflow in a NASH clinical trial setting with non-inferiority target for all assessment scores (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis) against a panel of qualified pathologists. If noninferiority is shown, then accuracy is also tested for superiority.
Secondary To describe the performance of pathologists assisted by AIM-NASH in determination of following aggregate NASH component scores: ● Clinical Research Network (CRN) fibrosis stage 4 (F4) vs other ● CRN fibrosis stage 0&1 (F0&1) vs other ● NAS aggregate score >4 vs other (NAS aggregate score here defined as the sum of ordinal scores for ballooning, steatosis, and inflammation).	**Secondary** Accuracy of aggregate NASH histologic component scores by computing concordance between AI-assisted and GT vs IMR and GT for the following aggregate component scores: ● CRN fibrosis stage F4 vs other ● CRN Fibrosis stages F0&F1 vs other ● NAS aggregate score >4 vs other (NAS aggregate score here defined as the sum of ordinal scores for ballooning, steatosis, and inflammation).
Exploratory: To describe overall performance of pathologists assisted by AIM-NASH in assessment of NASH at the component level in clinical subgroups (where available).	**Exploratory:** Overall accuracy analyses of pathologist reviewed AIM-NASH, per score component, in clinical groups defined by (where available and relevant) trial of origin, timepoint, and NASH treatment.

Additional post-hoc analyses to further examine performance in relevant inclusion subpopulations and/or aspects of efficacy endpoints included in results section

4.7.3 Study Design and Plan

PathAI utilized existing de-identified WSIs from partners from their completed clinical trials (screen failures and enrolled population from Intercept Pharmaceuticals REGENERATE trial NCT02548351, enrolled population from Bristol-Myers Squibb FALCON 2 trial NCT03486912 and enrolled population from Novo Nordisk Semaglutide trial NCT02970942). Each case utilized in this study is comprised of 2 slides – one H&E slide and

one trichrome slide. Slides were previously scanned for the clinical trial of origin by the sponsor's central or local laboratories using Aperio AT2 whole slide scanner at 40x magnification.

Each case enrolled into CV has a defined GT score, a minimum of 3 reads from the IMRs and one final AI-assisted read, unless otherwise noted (some cases had multiple AI-assisted reads for exploratory analyses and to ensure a wide range of cases was read by each pathologist). The GT and IMRs were performed on AISight Translational platform and for the AI-assisted, the AIM-NASH output was accessed on AISight Clinical Trials platform and the data was entered into the OpenClinica electronic data capture (eDC) platform.

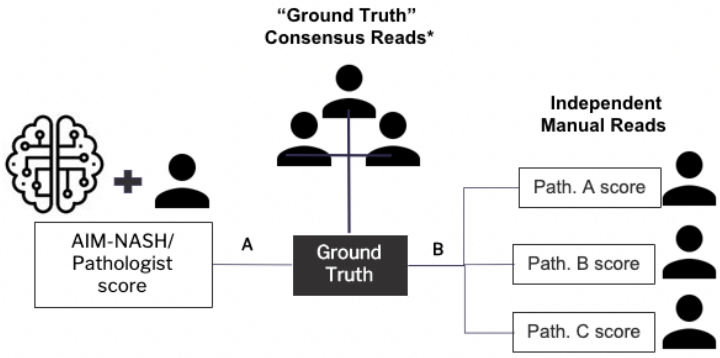
The GT was performed by two panels of 2 expert liver pathologists with a third pathologist serving as tiebreaker (Figure 41). The pathologists were chosen based on their previous experience and results of previously completed proficiency testing for PathAI (see Appendix IIa for the pathologist qualification and proficiency). All slides from analytical validation (AV) and CV were split between the two panels, so that each panel read about half of the overall dataset (AV and CV slides combined).

In cases where the two primary readers disagreed with the score on any of the NASH components (steatosis, lobular inflammation, hepatocellular ballooning and/or fibrosis), the slide was sent out to the third tiebreaker pathologist. To blind the tiebreaker to the discordant component, all NASH components were sent out for scoring. The tiebreaker pathologist was blinded to the scores of the two primary pathologists. If the tiebreaker pathologist agreed with one of the primary pathologists, this was then the final score for that NASH component. If the third pathologist disagreed with both primary pathologists, a joint panel call was held with the three pathologists to come to a consensus, with the tiebreaker providing the final score in the rare case that consensus was not reached. The tiebreaker pathologist was the same for both panels. Overall, 5 pathologists provided scores for GT. These 5 pathologists were unique and not used in any AI-assisted or for IMRs.

IMRs were performed by 8 qualified PathAI Contributor Network liver pathologists. Each CV slide was read by a minimum of 3 pathologists. Pathologists were selected by their previous experience in NASH trials and/or clinical experience with NASH, as well as performance tested to ensure proficiency (see Appendix I for pathologist qualification and proficiency). IMRs who scored the same slide did not have to come to a consensus, so no panel calls were held. For IMRs, not all slide pairs (H&E and trichrome) for each case were read by the same pathologist.

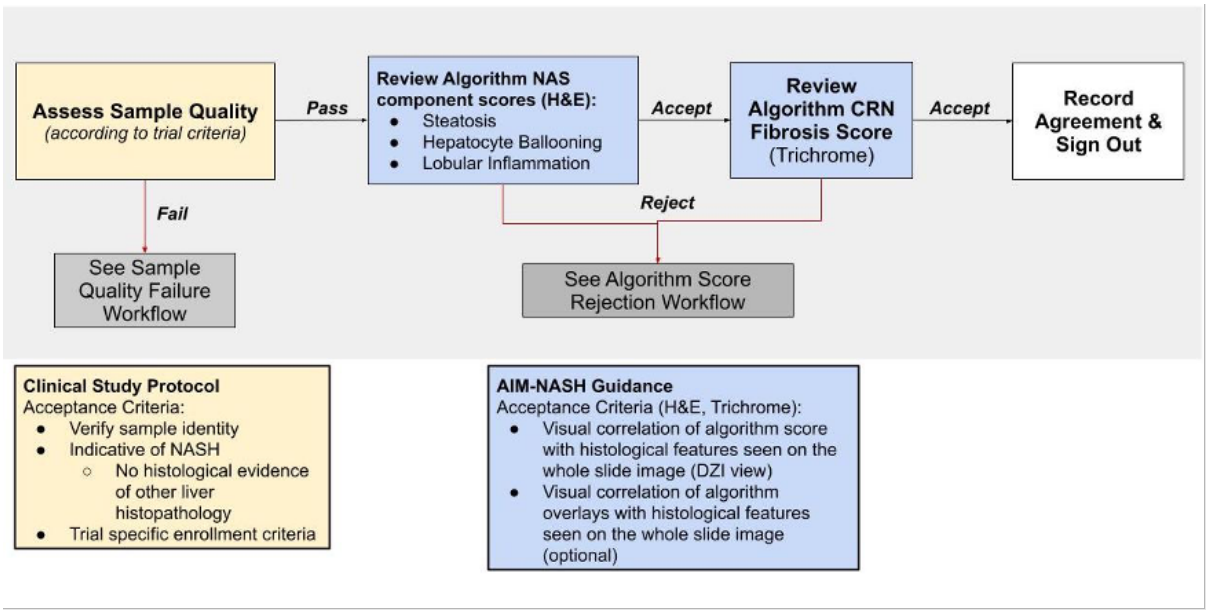
The AI-assisted reads were performed by a set of 6 expert liver pathologists. A subset of the IMRs were the same pathologists performing AI-assisted (Figure 42).

Figure 41: Clinical validation study design



The pathologist workflow for reviewing AIM-NASH results involves sample quality assessment, review of individual NAS component AIM-NASH scores on the H&E slide and review of the fibrosis algorithm score on the trichrome slide. The pathologist may choose to agree with all results and release the score or take actions in accordance with sample quality failure or follow the AIM-NASH result rejection workflow (described below). At each of these steps, the pathologist makes decisions in accordance with the AIM-NASH guidelines and clinical study protocol.

Figure 42: AI-assisted workflow



AIM-NASH workflow (AI-assisted):

Step A: Assess sample quality.

- Pathologist reviews the sample adequacy for acceptance:
    - Ensure sample is liver tissue. If there is non-liver tissue present, the pathologist is instructed to use the region exclusion tool to annotate the non-liver tissue and submit the case for AIM-NASH re-run.
    - Ensure that scanning artifact does not affect viewing of histologic features or prevent accurate scoring.
    - Ensure staining quality is adequate for scoring.
    - Ensure sample is adequate for scoring (sufficient number of portal tracts present).
    - Confirm histologic presence of NASH per protocol.
    - If there is histologic evidence of other liver histopathology, address per trial protocol.
    - Confirm trial enrollment criteria pertaining to the above are met.

- Acceptance criteria failed due to sample, stain or scan inadequacy (i.e., failed due to any of the criteria above) - Proceed to Step B
- Acceptance criteria passed (i.e., passed all the criteria above) - Proceed to Step C

Step B: Reject sample and request re-stain or rescan of the sample, as needed.

- For the purposes of the CV study, the pathologist indicates rejection of the sample but no re-stain or rescan activities were performed. In these instances, the pathologist finalized the case on the AISight Clinical Trials platform and indicated it was not evaluable in OpenClinica. If they were able to review the opposite stain, they were indicated to do so in OpenClinica.

Step C: Review AIM-NASH score for individual NAS components on a H&E slide.

- Inspect the WSI via the deep zoom image (DZI) view to ensure visual correlation of AIM-NASH scores with histologic features (steatosis, hepatocyte ballooning and lobular inflammation)
- Inspect the WSI via the DZI view to ensure visual correlation of AIM-NASH overlays with histologic features (steatosis, hepatocyte ballooning and lobular inflammation).
- If the pathologist disagrees with any of the AIM-NASH scores by a score of 2 or more, they reject the AIM-NASH score for that component (See Step F). If they agree or disagree with the score by 1 stage for all AIM-NASH scores, they will accept the score and move to the next step.

Step D: Review AIM-NASH score for fibrosis on a trichrome slide.

- Inspect the WSI via the DZI view to ensure visual correlation of AIM-NASH score with histologic features (fibrosis).
- Inspect the WSI via the DZI view to ensure visual correlation of AIM-NASH overlays with histologic features (fibrosis).
- If the pathologist disagrees with the AIM-NASH fibrosis score by a score of 2 or more, they reject the AIM-NASH score (See Step F). If they agree or disagree with the score by 1 stage, they will accept the score and move to the next step.

Step E: Record Agreement & Release Score

- If the pathologist agrees with all 4 NASH components (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis) then they accept the AIM-NASH results.

Step F: if the pathologist rejects any of the NASH components by a score of 2 or more, the case gets assigned to a secondary pathologist for a review.

Step G: the secondary pathologist independently reviews the discordant AIM-NASH score(s). The secondary pathologist is blinded to the primary pathologist's modified scores.

- For the purpose of the CV study, the secondary pathologist reviewed all AIM-NASH scores, however only the discordant components were included in downstream workflow and in the analysis.
- If the secondary pathologist's score agrees with the modified score provided by the primary pathologist, this is the final AI-assisted score.
- If the secondary pathologist's score disagrees with the modified score provided by the primary pathologist, a panel call is held.

Step H: Primary and secondary pathologists perform a consensus review.

- Pathologists hold a consensus call to discuss their independent assessments of AIM-NASH score(s). Only the NASH component(s) where the primary pathologist had a 2-stage disagreement are reviewed on the call.

Step I: Primary and secondary pathologists either agree with the AIM-NASH score or provide a modified score that is at least 2 or more stages different from the AIM-NASH score.

4.7.4 Dataset

Table 81: Clinical Validation Population

Trial Name and Sponsor	Trial Phase	Drug/Drug Class	Enrollment Criteria	Total Available Sample Size from Original Trial
REGENERATE Intercept Pharmaceuticals	3	Obeticholic Acid/ Farnesoid X receptor agonist	Presence of all 3 NAS components Fibrosis stage 2 or stage 3 OR Fibrosis stage 1a or stage 1b if accompanied by ≥1 of the following risk factors: Obesity (BMI ≥30 kg/m2) Type 2 diabetes diagnosed per 2013 American Diabetes Association criteria ALT >1.5× upper limit of normal (ULN).	5000
FALCON2 Bristol Myers Squibb	2	Pegbelfermin / Pegylated FGF21 (anti-fibrotic)	Biopsy must be consistent with NASH Biopsy must be consistent with cirrhosis (stage 4)	281
Semaglutide NASH Trial (NCT04822181) Novo Nordisk	2	Semaglutide / GLP-1 Agonist	Biopsy-proven NASH; A histological NAFLD activity score equal to or above 4 with a score of 1 or more in steatosis, lobular inflammation, and hepatocyte ballooning Fibrosis stage 2,3	526

4.7.5 Selection of Study Population/ Cases

Inclusion Criteria

- De-identified NASH biopsy slides (H&E and trichrome-stained slides from two phase 2 (one F4) and one phase 3 NASH clinical trials) from screening and intermediate analysis (IA) populations, representing a wide spectrum of disease (characterized by NAS grade and CRN fibrosis stage), a variety of staining (performed in different labs), multiple time points of collection (baseline, post baseline), and from trials with different classes of drug targets.

Exclusion Criteria

- Slides that are non-evaluable by manual read will be excluded.
- Only trichrome blue will be included. Any other trichrome protocols (e.g., green counterstain) will be excluded from this validation and context of use.

4.7.6 General Procedures

Blinding

All participating pathologists had their own unique log in to the AISight Translational platform, AISight Clinical Trials platform and/or OpenClinica eDC. Each pathologist was assigned cases with relevant slides. All study

pathologists were blinded to each other's individual scores, as well as any trial, timepoint, or treatment arm information. IMRs and GT pathologists were not presented with AIM-NASH scores. All PathAI staff (except for the unblinded clinical data managers and unblinded clinical scientists) involved in this study were blinded to the data until the database was locked.

Glass Slides Scanning and Handling
Not applicable, no glass slides were utilized in the study. All slides were previously scanned using the Aperio AT2 scanner at 40x in CLIA/CAP certified laboratories following laboratory Standard Operating Procedures (SOPs).

4.7.7 Pathologist Training

All pathologists have experience in reading for NASH clinical trials and/or signing out potential NASH cases clinically. For this study, the AI-assisted NASH pathologists participated in a training webinar where the use of AISight Clinical Trials platform and OpenClinica was demonstrated, as well as examples of AIM-NASH outputs were discussed. Attendance was recorded in PathAI's eQMS system.

All GT and IMRs pathologists received and were asked to review CRN NASH histology scoring guidelines and corresponding literature prior to completing any NASH scoring tasks. They received separate instructions for each H&E and trichrome scoring task. These instructions detailed how many slides were to be reviewed and the type of review needed. These instructions also included the CRN NASH scoring criteria.

4.7.8 Data Handling

All GT and IMR data were entered electronically on the AISight Translational platform, and all AI-assisted data was entered in the OpenClinica eDC platform. AIM-NASH algorithm was run on the AISight Clinical Trials platform, and the scores (without pathologist review) were securely downloaded for data querying purposes. After the completion of the study, all data was securely downloaded from the platforms and stored in the clinical data management's Amazon Web Services (AWS) bucket. Data for analysis was uploaded to the PathAI's eQMS system after the database lock. PathAI designated clinical data managers and clinical scientists who were unblinded to the data and had access to all study information for the purpose of monitoring data and resolving any queries.

Data Quality Assurance
PathAI utilizes SOPs designed to ensure that research procedures and documentation are consistently conducted/prepared to the highest quality standards. These SOPs also require compliance with ICH Guideline for Good Clinical Practice (E6 (4.6.4)).

4.7.9 Statistical Methods and Determination of Sample Size

Primary Analysis
Accuracy of the AI-assisted scores are assessed separately for each of the 4 NASH component scores (steatosis, lobular inflammation, hepatocellular ballooning and fibrosis). A single AI-assisted score, a single GT score, and a set of at least 3 IMR scores are collected for each NASH histologic component. As a reference, mean (across IMRs) pathologist-GT concordance will be quantified by the linearly weighted Kappa (WK) statistic for each NASH component (represented by "B" in

Table 39; for expected values, see Table 40). AI-assisted agreement with GT will be quantified by linearly WK between AI-assisted and GT for each NASH component (represented by "A" in
Table 39). To show agreement non-inferior to a standard qualified pathologist, AI-assisted concordance with GT pathologists will be shown to be significantly greater (Bootstrap percentile $p < 0.025$) than 0.1 less than the mean IMR concordance with GT for each NASH assessment (H1: $A > (B - 0.1)$). Overall acceptance will be determined by meeting this non-inferiority criteria for all 4 NASH assessments (steatosis, ballooning, lobular inflammation, and fibrosis). If non-inferiority is shown, then accuracy will also be tested for superiority of AI-assisted concordance with GT to IMR concordance with the GT.

Secondary analyses

Accuracy of aggregate NASH component scores will be assessed by computing concordance between AI-assisted and GT and each IMR and GT for the following aggregate component scores:

- Fibrosis stage 4 vs other
- Fibrosis stage 0 and 1 vs other
- NAS aggregate score ≥4 vs other (NAS aggregate score here defined as the sum of component scores for steatosis, lobular inflammation, and hepatocellular ballooning).

Linearly WK concordance values for these comparisons will be presented separately for AIM-NASH vs GT and the average of each IMR vs GT. 95% confidence intervals will be obtained from the bootstrap percentile method.

4.7.10 CV Results

Data Sets Analyzed

Each case included in the study has a minimum of 1 AI-assisted score, minimum of 2 initial scores from GT pathologists, as well as a final consensus score where applicable, and a set of at least 3 IMR scores for each NASH component. All pathologists were blinded to timepoint or any original trial information. Any slide where at least 2 GT pathologists indicated the biopsy was not adequate for scoring was removed from the analysis. Any NASH histologic component (steatosis, lobular inflammation, hepatocellular ballooning and/or fibrosis) where at least 2 GT pathologists determined the slide to be not evaluable for that component, was removed from the analysis, even if AI-assisted scores and/or IMR scores were collected for these slides. For any NASH histologic components (steatosis, lobular inflammation, hepatocellular ballooning and/or fibrosis) where the GT deemed the slide adequate for the NASH score component, but an individual manual pathologist deemed it inadequate, that score component was removed from analysis for that individual manual pathologist only. In cases where the AI-assisted workflow deemed the slide inadequate, the slide was removed from analysis for AI-assisted workflow only.

In order to ensure that all AI-assisted workflow pathologists read cases with a variety of scores, some cases were sent out to more than one pathologist for AIM-NASH review. These cases were defined as "spike-in" cases. All spike-in cases were excluded from the accuracy analyses and only the original base cases were included. For GT slides that were inadvertently sent out to a tiebreaker when the 2 primary pathologists had agreed on the score or sent to a consensus call when the score had been decided in tiebreaker. For these cases, the majority score was used in analysis instead of the consensus score. Additionally, for cases where one of the GT primary pathologists was missing a score for a NASH component or gave a non-standard answer (e.g., not applicable for a score component when the slide was deemed evaluable by that pathologist), the slide was sent to consensus before it

was queried, in these cases, the consensus score was used for analysis. For cases where the case was sent to a consensus call when it should have been sent to a tiebreaker, the consensus call scores were used in analysis. If any slides were sent out more than once, the scores from the job sent out the earliest were used in the analysis.

For AI-assisted workflow, in some instances, the same case was inadvertently assigned to more than one AI-assisted workflow pathologist. For these cases, scores from the first pathologist assigned to the case were used for analysis, the other pathologist(s) scores were removed from the analysis. Additionally, for some cases, a pathologist who was not assigned to the case reviewed the case and created a new CRF in OpenClinica; in these cases, the data was removed from analysis and data from the original pathologist assigned to the case was used. Any cases for AI-assisted workflow, where the case was sent to secondary review and/or panel call before initial reviewer answers were queried, the secondary review/panel call data was removed from the analysis if the resolved query eliminated the need for a secondary review and/or panel call. In instances where a case had incorrect images uploaded for AI-assisted workflow, the data collected from that case was removed from analysis.

For all cases where scores were collected for trichrome green slides or for slides that were out of focus, these scores were removed from analysis. For any slides where the incorrect CRF was sent out (e.g., H&E CRF for a trichrome slide), the data from these CRFs was also excluded from analysis.

Two cases were removed from data analysis because the patient from the original clinical trial did not consent to further research.
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%) (Table 82).

Table 82: Reason for Missing Final GT Score

Component	Reason	% (n/N)
Steatosis	Other/Reason not provided, Sample	0.07 (1/1501)
	Sample	0.87 (13/1501)
	Sample, Scan	0.33 (5/1501)
	Sample, Stain	0.07 (1/1501)
Lobular inflammation	Other/Reason not provided	0.13 (2/1501)
	Other/Reason not provided, Sample	0.07 (1/1501)
	Sample	0.93 (14/1501)
	Sample, Scan	0.33 (5/1501)
	Sample, Stain	0.07 (1/1501)
Hepatocellular ballooning	Other/Reason not provided, Sample	0.07 (1/1501)
	Sample	1.07 (16/1501)
	Sample, Scan	0.33 (5/1501)
	Sample, Stain	0.07 (1/1501)
	Scan	0.07 (1/1501)
	Stain	0.07 (1/1501)
Fibrosis	Other/Reason not provided	0.07 (1/1501)
	Other/Reason not provided, Sample	0.13 (2/1501)
	Other/Reason not provided, Sample, Stain	0.13 (2/1501)
	Sample	1.4 (21/1501)
	Sample, Scan	0.07 (1/1501)
	Sample, Stain	0.87 (13/1501)
	Stain	0.53 (8/1501)

Similarly, less than 1% of the slides had a missing score from all IMR pathologists reviewing the slide for all components (

Table 83).

Table 83: Reason for Missing IMR Score

Component	Reason	% (n/N)
Steatosis	Scan	0.07 (1/1501)
Lobular inflammation	Sample	0.07 (1/1501)
	Scan	0.13 (2/1501)
Hepatocellular ballooning	Sample	0.07 (1/1501)
	Scan	0.2 (3/1501)
Fibrosis	Sample	0.13 (2/1501)
	Stain	0.07 (1/1501)

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 (Table 84). There were 7 slides where AIM-NASH was not able to provide a score due to blurry images.

Table 84: Reason for Missing AI-assisted Score due to Pathologist

Component	Reason	% (n/N)
Steatosis	Sample	1.4 (21/1501)
	Scan	0.4 (6/1501)
Lobular inflammation	Sample	1.4 (21/1501)
	Scan	0.4 (6/1501)
Hepatocellular ballooning	Sample	1.4 (21/1501)
	Scan	0.4 (6/1501)
Fibrosis	Sample	2.1 (32/1501)
	Scan	0.2 (3/1501)
	Stain	0.9 (14/1501)

Study Population

No demographic information for the slides enrolled in the study is available. The dataset represents both screen-failed and enrolled NASH clinical trial patient populations, including study subjects who may have regressed or progressed during a clinical trial, and reflects the NASH patient population as a whole. The dataset also contains variability in sample staining and scanning (including performed by multiple collection/preparation sites and central laboratories). Distribution of slides based on slide level score from glass GT are listed in

Table 85.

Table 85: Slide distribution by final GT score

Feature	Score	% (n/N)
Steatosis	0	8.37 (124/1481)
	1	43.48 (644/1481)
	2	34.5 (511/1481)
	3	13.64 (202/1481)
Lobular inflammation	0	0.81 (12/1478)
	1	63.46 (938/1478)
	2	33.69 (498/1478)
	3	2.03 (30/1478)
Hepatocellular ballooning	0	11.11 (164/1476)
	1	47.02 (694/1476)
	2	41.87 (618/1476)
Fibrosis	0	0.89 (13/1453)
	1	12.04 (175/1453)
	2	27.25 (396/1453)
	3	38.89 (565/1453)
	4	20.92 (304/1453)
NAS	0	0.54 (8/1474)
	1	4.61 (68/1474)
	2	6.72 (99/1474)
	3	18.66 (275/1474)
	4	25.03 (369/1474)
	5	24.08 (355/1474)
	6	15.94 (235/1474)
	7	4.14 (61/1474)
	8	0.27 (4/1474)

4.7.11 Results

Primary endpoint

Non-inferior accuracy was evaluated for AI-assisted reads compared to unassisted pathologist reads (IMRs) for all NASH histologic components (steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis), using a panel of qualified pathologists for GT. If noninferiority was achieved, then accuracy was also evaluated for superiority. For all 4 NASH components, non-inferiority was achieved (Figure 43, Figure 44 and Table 86). For hepatocellular ballooning and lobular inflammation, superiority was also achieved for AI-assisted versus unassisted reads (Figure 43, Figure 44 and Table 86). The difference in agreement for AI-assisted and GT compared to agreement for mean IMR and GT for hepatocellular ballooning was 0.15 (95% CI of (0.108, 0.195); NI p<0.0001) and for lobular inflammation was 0.123 (95% CI of (0.069, 0.173); NI p<0.0001) with a p<0.0001 for superiority for both components. The difference in agreement for AI-assisted and GT compared to agreement of mean IMR and GT for steatosis was 0.003 (95% CI of (-0.028, 0.037); NI p<0.0001) and for fibrosis was 0.008 (95% CI of (-0.026, 0.039); NI p<0.0001). Steatosis and fibrosis did not achieve superiority but were well within the non-inferiority margin. For all NASH score components the WKs for AI-assisted and GT were in the ranges of published CRN pathologists WKs (Table 40, Figure 44, Table 86) (14,22).

Figure 43: Accuracy results for each NASH histologic component

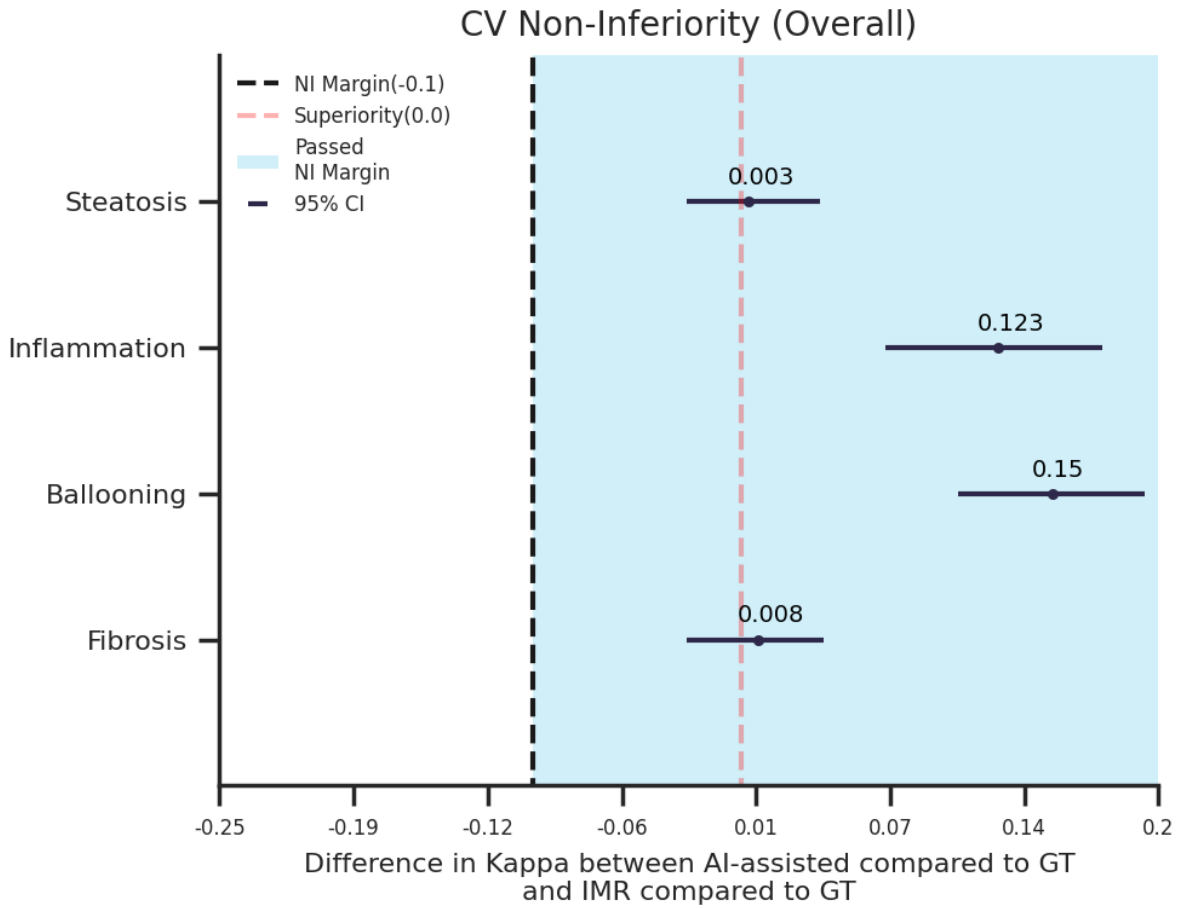


Figure 44: Accuracy concordance comparison of NASH histologic components

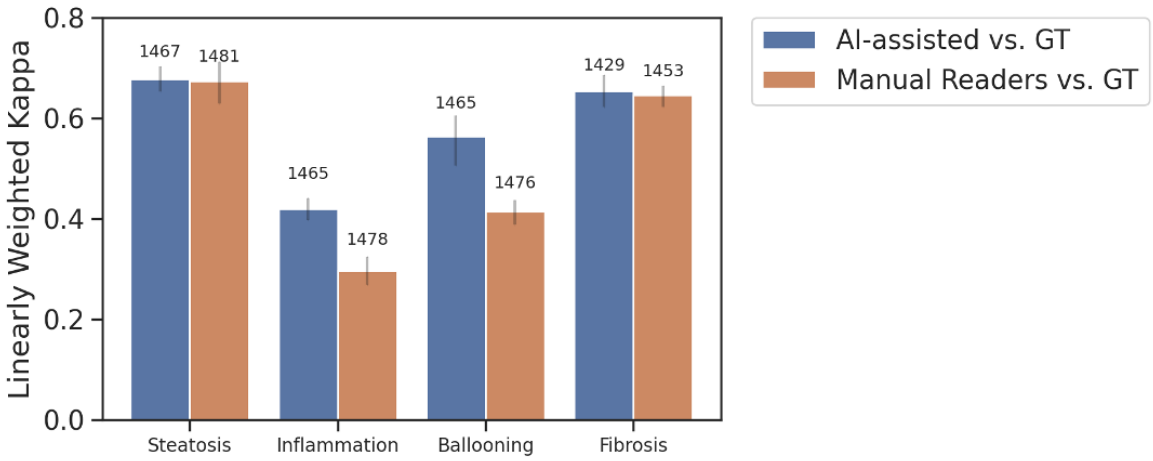


Table 86: Primary endpoint results for each NASH histologic component

Feature	Modality	N	WK (95% CI)	Difference (95% CI)	p-value for NI	p-value for Superiority
Steatosis	AI-assisted vs GT	1467	0.677 (0.652, 0.703)	0.003 (-0.026, 0.038)	<0.0001	0.3455
	IMR vs GT	1481	0.674 (0.651, 0.694)			
Lobular inflammation	AI-assisted vs GT	1465	0.419 (0.361, 0.46)	0.123 (0.069, 0.173)	<0.0001	<0.0001
	IMR vs GT	1478	0.297 (0.265, 0.329)			
Hepatocellular ballooning	AI-assisted vs GT	1465	0.563 (0.519, 0.601)	0.15 (0.104, 0.194)	<0.0001	<0.0001
	IMR vs GT	1476	0.414 (0.385, 0.442)			
Fibrosis	AI-assisted vs GT	1429	0.653 (0.627, 0.676)	0.008 (-0.026, 0.039)	<0.0001	0.42
	IMR vs GT	1453	0.645 (0.622, 0.665)			

Secondary endpoint

For the secondary endpoint, WKs for aggregate NASH component scores (F0&F1 vs other, F4 vs other and NAS $\geq$ 4 vs NAS < 4) were computed between AI-assisted and GT and between IMR and GT (Figure 45 and Table 87). The WKs between AI-assisted and GT and between IMR and GT for all 3 aggregate NASH component scores were similar with overlapping confidence intervals, with F4 vs other and NAS $\geq$ 4 vs other having a higher WK

for AI-assisted and GT compared to IMR and GT (0.753 vs 0.705 for F4 vs other and 0.674 vs 0.577 for NAS ≥ 4 vs other). For F0&F1 vs other, the WK for IMR and GT was higher than AI-assisted and GT (0.539 vs 0.497, respectively).

Figure 45: WKs for NASH aggregate scores

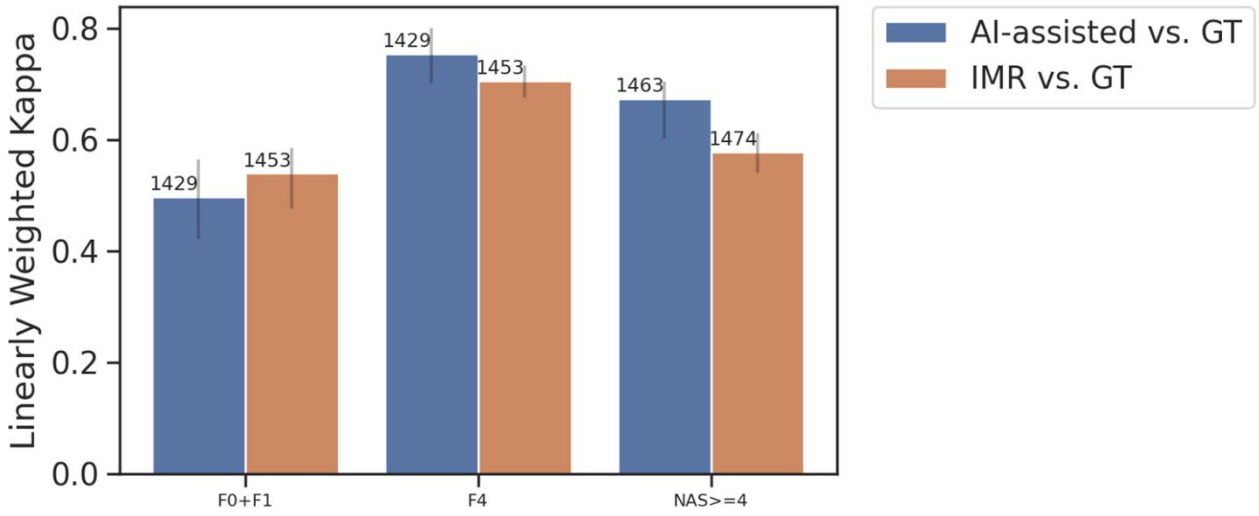


Table 87: WKs for aggregate NASH component scores

Aggregate Score	Modality	N	WK (95% CI)	Difference (95% CI)
F0&F1 vs other	AI-assisted vs GT	1429	0.497 (0.421, 0.563)	-0.042 (-0.129, 0.036)
	IMR vs GT	1453	0.539 (0.488, 0.585)	
F4 vs other	AI-assisted vs GT	1429	0.753 (0.683, 0.784)	0.048 (-0.026, 0.121)
	IMR vs GT	1453	0.705 (0.642, 0.750)	
NAS ≥4 vs. <4	AI-assisted vs GT	1463	0.674 (0.645, 0.701)	0.097 (0.048, 0.142)
	IMR vs GT	1474	0.577 (0.542, 0.610)	

Exploratory endpoints

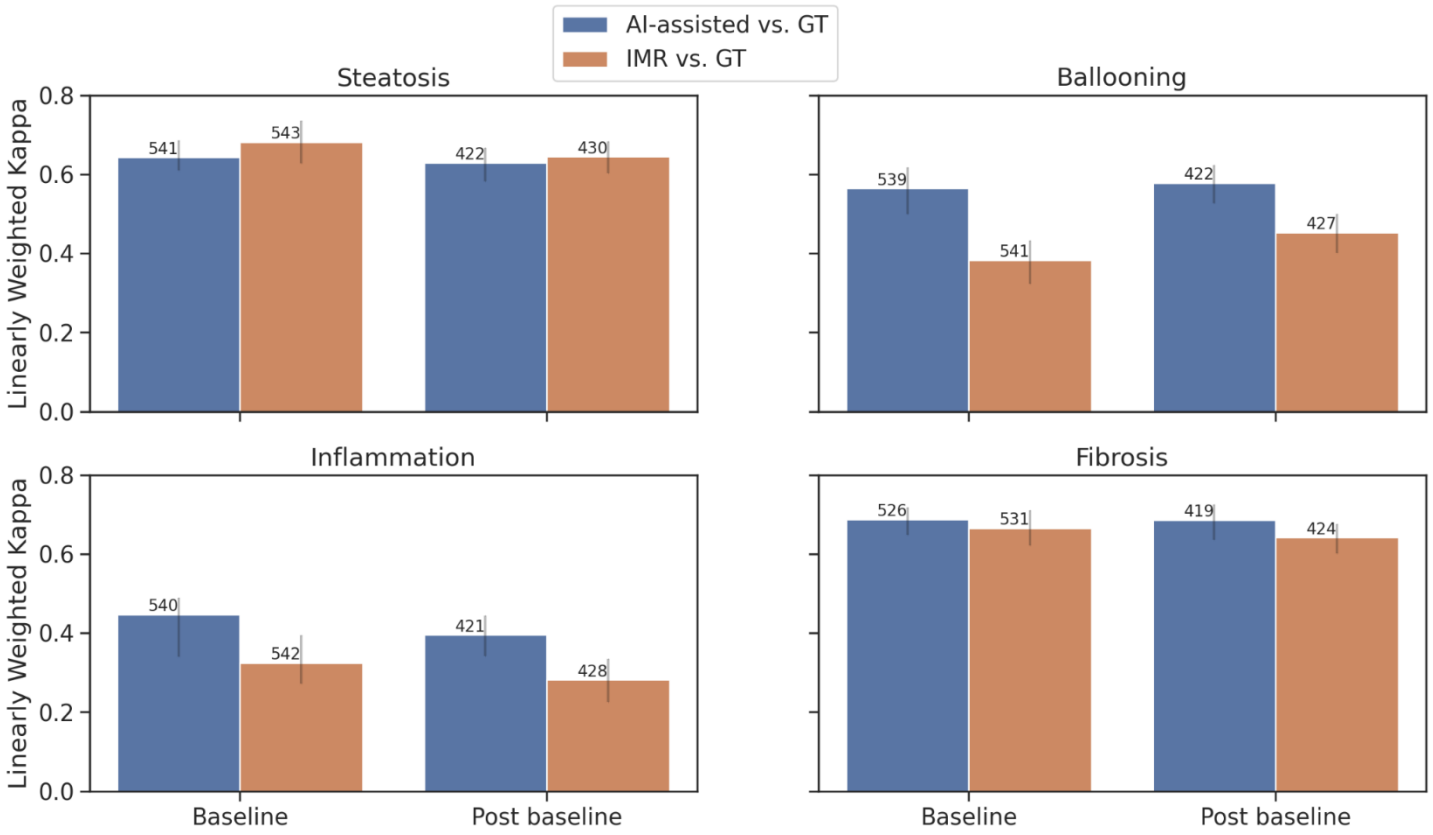
Predefined exploratory endpoints

The predefined exploratory endpoints explored the accuracy of AI-assisted scores in clinical trial relevant subsets by time point, sponsor, and score level components, where available. Additionally, the utility of NAS (hepatocellular ballooning, steatosis, and lobular inflammation) and fibrosis overlays was assessed.

- **Agreement by timepoint**

Time point information was available for samples from FALCON 2 and REGENERATE clinical trials but was not available for Novo Nordisk Semaglutide trial and therefore Novo Nordisk samples are not included in this analysis. For FALCON 2 and REGENERATE steatosis, WKs for AI-assisted and GT were similar to WKs for IMR and GT, with IMR and GT WKs being slightly higher for both baseline and post baseline samples (Figure 46 and Table 88). For fibrosis, WKs were very similar. However, for hepatocellular ballooning and lobular inflammation, the WKs for AI-assisted and GT were substantially higher than IMR and GT for both baseline and post baseline time points (Figure 46 and Table 88)

*Figure 46: WKs per NASH component per time point for trials with available timepoint data (Falcon 2 and Regenerate *).*



*Timepoint data from the Novo Nordisk Semaglutide ph2b trial was not available

Table 88: WKs for NASH components per time point for Falcon 2 and Regenerate

Feature	Visit	Modality	N	WK (95% CI)	Difference (95% CI)
Steatosis	Baseline	AI-assisted vs GT	541	0.643 (0.611, 0.686)	-0.039 (-0.092, 0.023)
		IMR vs GT	543	0.682 (0.636, 0.720)	
	Post baseline	AI-assisted vs GT	422	0.629 (0.575, 0.682)	-0.015 (-0.080, 0.052)
		IMR vs GT	430	0.645 (0.602, 0.682)	
Lobular inflammation	Baseline	AI-assisted vs GT	540	0.447 (0.34, 0.489)	0.122 (0.018, 0.190)
		IMR vs GT	542	0.325 (0.271, 0.373)	
	Post baseline	AI-assisted vs GT	421	0.396 (0.343, 0.465)	0.114 (0.038, 0.205)
		IMR vs GT	428	0.282 (0.225, 0.335)	
Hepatocellular ballooning	Baseline	AI-assisted vs GT	539	0.565 (0.499, 0.617)	0.181 (0.106, 0.255)
		IMR vs GT	541	0.383 (0.331, 0.428)	
	Post baseline	AI-assisted vs GT	422	0.578 (0.517, 0.628)	0.125 (0.055, 0.188)
		IMR vs GT	427	0.453 (0.402, 0.500)	
Fibrosis	Baseline	AI-assisted vs GT	526	0.687 (0.648, 0.717)	0.022 (-0.032, 0.082)
		IMR vs GT	531	0.665 (0.616, 0.704)	
	Post baseline	AI-assisted vs GT	419	0.686 (0.642, 0.732)	0.045(-0.016, 0.106)
		IMR vs GT	424	0.641 (0.602, 0.675)	

- **Agreement by Sponsor**

Exploratory analysis per sponsor (trial of origin) for each NASH histologic component is shown in Figure 47 and Table 89. For steatosis and fibrosis, by sponsor, the WKs for AI-assisted and GT were similar to WKs for IMR and GT, with overlapping confidence intervals. For lobular inflammation and hepatocellular ballooning, the WKs for AI-assisted and GT were higher than the WKs for IMR and GT for all 3 sponsors. For lobular inflammation, from Intercept's REGENERATE and for Novo Nordisk's Semaglutide trial, the WKs for AI-assisted reads were significantly higher than WK for IMR with a difference of 0.168 (95% CI of (0.094, 0.230)) and 0.126 (95% CI of (0.047, 0.213)), respectively. Similarly, for hepatocellular ballooning, the WKs for AI-assisted reads were significantly higher than WKs for IMR with a difference of 0.176 (95% CI of (0.118, 0.233)) and 0.133 (95% CI of (0.019, 0.232)), respectively.

Figure 47: WKs per NASH component per trial sponsor.

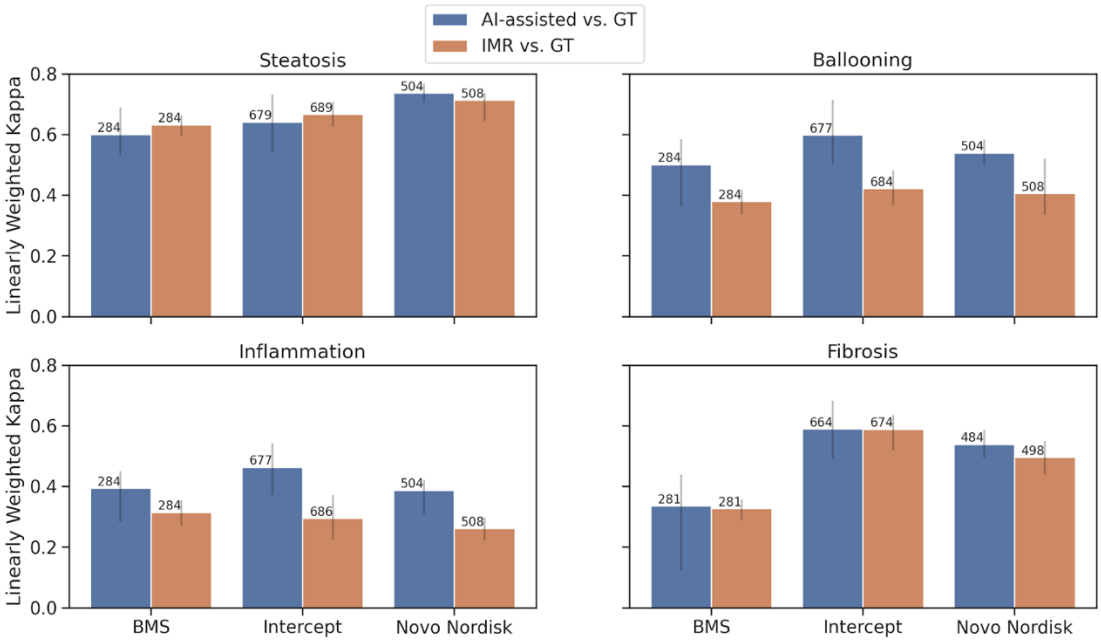


Table 89: WKs for NASH components per sponsor

Feature	Sponsor	Modality	N	WK (95% CI)	Difference (95% CI)
Steatosis	BMS	AI-assisted vs GT	284	0.600 (0.532,0.689)	-0.032 (-0.150,0.089)
		IMR vs GT	284	0.632 (0.535,0.723)	
	Intercept	AI-assisted vs GT	679	0.641 (0.611/0.672)	-0.025 (-0.064,0.019)
		IMR vs GT	689	0.666 (0.632,0.697)	
	Novo Nordisk	AI-assisted vs GT	504	0.736 (0.696, 0.777)	0.023 (-0.023,0.110)
		IMR vs GT	508	0.713 (0.645, 0.737)	
Lobular inflammation	BMS	AI-assisted vs GT	284	0.394 (0.281, 0.450)	0.079 (-0.042,0.176)
		IMR vs GT	284	0.315 (0.224, 0.393)	
	Intercept	AI-assisted vs GT	677	0.463 (0.385, 0.497)	0.168 (0.094,0.230)
		IMR vs GT	686	0.295 (0.251, 0.334)	
	Novo Nordisk	AI-assisted vs GT	504	0.387 (0.318, 0.464)	0.126 (0.047,0.213)
		IMR vs GT	508	0.261 (0.224, 0.297)	

Hepatocellular ballooning	BMS	AI-assisted vs GT	284	0.501 (0.363, 0.584)	0.122 (-0.055,0.258)
		IMR vs GT	284	0.379 (0.286, 0.495)	
	Intercept	AI-assisted vs GT	677	0.598 (0.558, 0.641)	0.176 (0.118,0.233)
		IMR vs GT	684	0.422 (0.380, 0.460)	
	Novo Nordisk	AI-assisted vs GT	504	0.539 (0.486, 0.598)	0.133 (0.019,0.232)
		IMR vs GT	508	0.406 (0.336, 0.520)	
Fibrosis	BMS	AI-assisted vs GT	281	0.335 (0.124, 0.438)	0.009 (-0.200,0.147)
		IMR vs GT	281	0.327 (0.229, 0.419)	
	Intercept	AI-assisted vs GT	664	0.589 (0.545, 0.635)	0 (-0.052, 0.055)
		IMR vs GT	674	0.588 (0.553, 0.619)	
	Novo Nordisk	AI-assisted vs GT	484	0.539 (0.47, 0.585)	0.043 (-0.04,0.117)
		IMR vs GT	498	0.496 (0.441, 0.548)	

- **Agreement by Score Level**

Exploratory analysis per individual score levels for each NASH histologic component are shown in Figure 48 and Table 90. WKs for AI-assisted and GT for hepatocellular ballooning were significantly higher for all scores (0, 1 and 2) than WKs for IMR and GT. For steatosis, WKs were largely similar with overlapping confidence intervals, except for steatosis scores of 2 and 3, where the WK for AI-assisted and GT was significantly higher than the average WK for IMR and GT. For lobular inflammation, the WKs for AI-assisted and GT were significantly higher for scores 1 and 2 than WKs for IMR and GT and equivalent for scores of 0 and 3, with overlapping confidence intervals. However, the number of reads for scores 0 and 3 were quite low (AI-assisted/GT n=7 and IMR/GT n=9 for score 0 and AI-assisted/GT n=5 and IMR/GT n=25 for score 3). For fibrosis, the WKs for AI-assisted and GT and the WKs for IMR and GT were largely the same, with overlapping confidence intervals for every score.

Figure 48: WKs per NASH component per score.

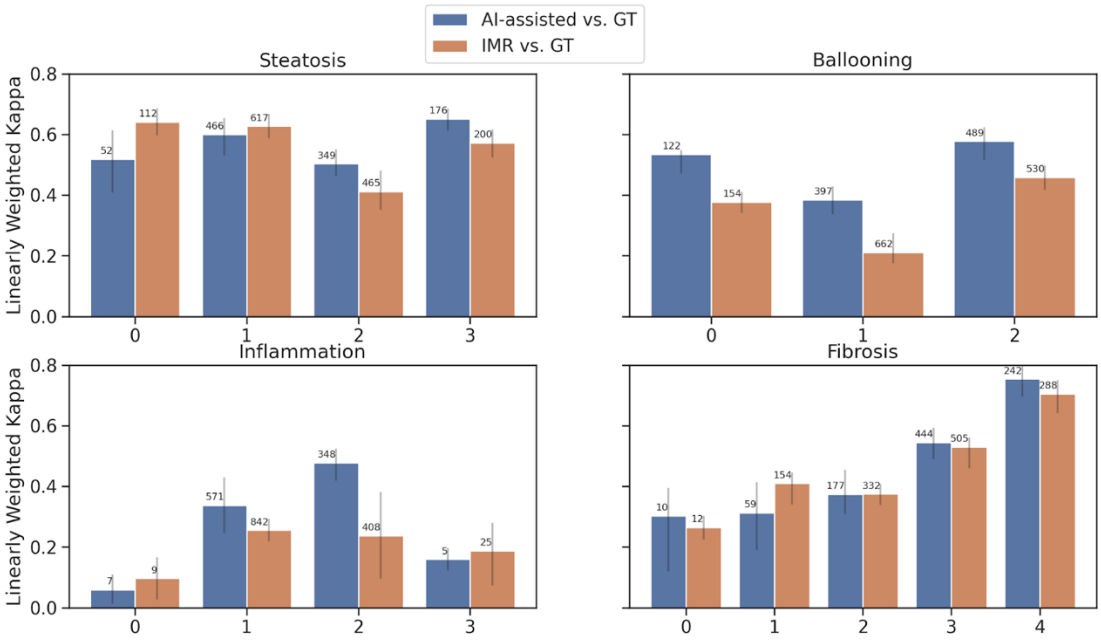


Table 90: WKs for NASH components per score

Feature	**Score**	**Modality**	**N**	**WK (95% CI)**	**Difference (95% CI)**
Steatosis	0	AI-assisted vs GT	52	0.519 (0.409, 0.613)	-0.122 (-0.245, -0.002)
		IMR vs GT	112	0.641 (0.571, 0.694)	
	1	AI-assisted vs GT	466	0.600 (0.561, 0.647)	-0.027 (-0.0774, 0.029)
		IMR vs GT	617	0.627 (0.589, 0.660)	
	2	AI-assisted vs GT	349	0.503 (0.459, 0.549)	0.091 (0.037, 0.150)
		IMR vs GT	465	0.412 (0.372, 0.451)	
	3	AI-assisted vs GT	176	0.651 (0.591, 0.72)	0.079 (0.009, 0.160)
		IMR vs GT	200	0.573 (0.524, 0.616)	
Lobular inflammation	0	AI-assisted vs GT	7	0.059 (0.014, 0.109)	-0.038 (-0.137, 0.066)
		IMR vs GT	9	0.097 (0.006, 0.188)	
	1	AI-assisted vs GT	571	0.338 (0.279, 0.384)	0.083 (0.021, 0.137)
		IMR vs GT	842	0.255 (0.220, 0.292)	

	2	AI-assisted vs GT	348	0.478 (0.408, 0.547)	0.24 (0.169, 0.311)
		IMR vs GT	408	0.237 (0.202, 0.274)	
	3	AI-assisted vs GT	5	0.159 (0.018, 0.303)	-0.028 (-0.191, 0.156)
		IMR vs GT	25	0.187 (0.074, 0.279)	
Hepatocellular ballooning	0	AI-assisted vs GT	122	0.535 (0.472, 0.548)	0.158 (0.081, 0.207)
		IMR vs GT	154	0.377 (0.330, 0.420)	
	1	AI-assisted vs GT	397	0.385 (0.324, 0.430)	0.174 (0.118, 0.225)
		IMR vs GT	662	0.211 (0.176, 0.244)	
	2	AI-assisted vs GT	489	0.578 (0.543, 0.640)	0.119 (0.070, 0.186)
		IMR vs GT	530	0.459 (0.417, 0.4987)	
Fibrosis	0	AI-assisted vs GT	10	0.303 (0.12, 0.394)	0.038 (-0.156, 0.195)
		IMR vs GT	12	0.265 (0.143, 0.365)	
	1	AI-assisted vs GT	59	0.313 (0.249, 0.393)	-0.097 (-0.181, -0.011)
		IMR vs GT	154	0.410 (0.356, 0.457)	
	2	AI-assisted vs GT	177	0.374 (0.316, 0.443)	-0.001 (-0.067, 0.073)
		IMR vs GT	332	0.375 (0.335, 0.412)	
	3	AI-assisted vs GT	444	0.544 (0.476, 0.580)	0.014 (-0.061, 0.064)
		IMR vs GT	505	0.530 (0.495, 0.563)	
	4	AI-assisted vs GT	242	0.753 (0.683, 0.784)	0.048 (-0.026, 0.121)
		IMR vs GT	288	0.705 (0.642, 0.750)	

- **Clinical Utility**

To explore the utility of the AIM-NASH overlays as a feature highlight for the pathologists in reviewing the AIM-NASH score, the pathologists were asked to rate the utility of the overlay on a scale of 1-5, with 1 being not useful at all and 5 being very useful. Utility rating of the initial pathologists as well as the secondary pathologists, where available, is considered for this analysis. (Figure 49 and Table 91). For steatosis, lobular inflammation and hepatocellular ballooning, the pathologists found that the overlays were not at all useful about 2% of the time. Fibrosis overlays were found to be not at all useful about 5% of the time. It's important to note that the pathologists may toggle the overlays on and off as they see fit and the use of overlays is not tracked or recorded.

Figure 49: Utility of AIM-NASH overlays for primary and secondary reviewers

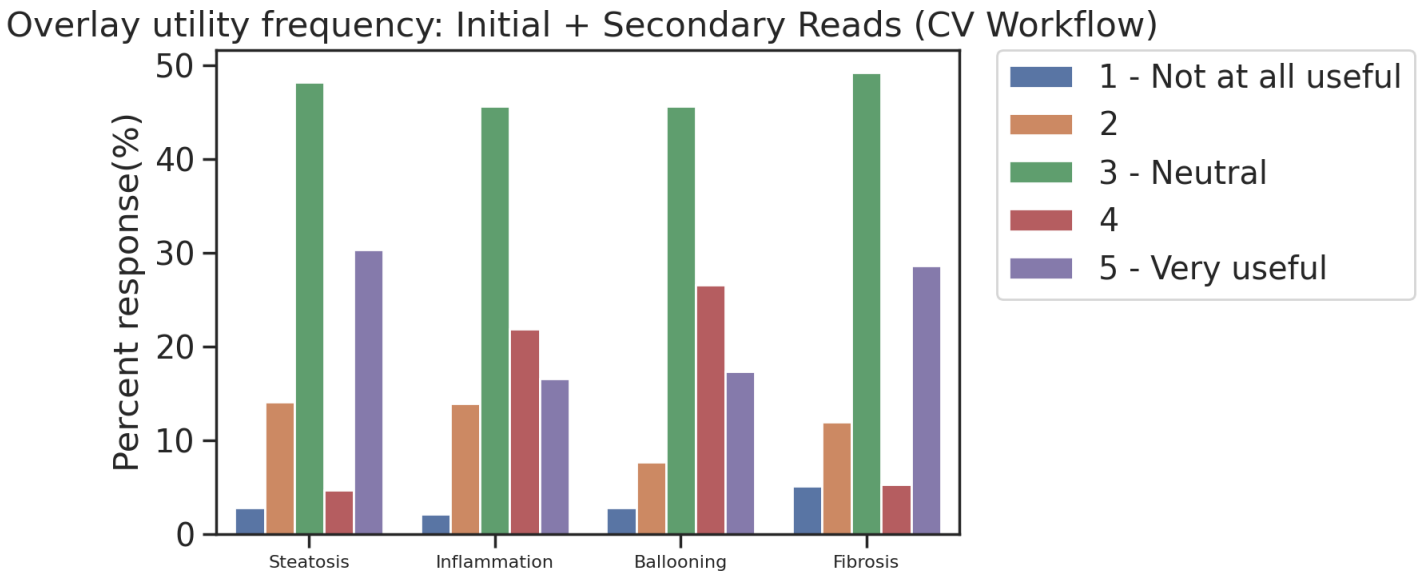


Table 91: Utility of AIM-NASH overlays for primary and secondary reviewers

Feature	**Overlay Utility**	**% (n/N)**
Steatosis	1 - Not at all useful	2.83 (48/1694)
	2	14.05 (238/1694)
	3 - Neutral	48.17 (816/1694)
	4	4.66 (79/1694)
	5 - Very useful	30.28 (513/1694)
Lobular inflammation	1 - Not at all useful	2.07 (35/1694)
	2	13.93 (236/1694)
	3 - Neutral	45.57 (772/1694)
	4	21.84 (370/1694)
	5 - Very useful	16.59 (281/1694)
Hepatocellular ballooning	1- Not at all useful	2.77 (47/1694)
	2	7.67 (130/1694)
	3- Neutral	45.63 (773/1694)
	4	26.56 (450/1694)

	5- Very useful	17.36 (294/1694)
Fibrosis	1 - Not at all useful	5.08 (86/1694)
	2	11.92 (202/1694)
	3 - Neutral	49.17 (833/1694)
	4	5.25 (89/1694)
	5 - Very useful	28.57 (484/1694)

Post-hoc Exploratory Endpoints

- **Agreement for clinical trial relevant aggregate component scores**

In post-hoc exploratory analysis, WKs for additional relevant trial inclusion criteria (F2&F3 vs other and NAS $\geq$ 4 with $\geq$1 in each score category vs other) were computed between AI-assisted and GT and between IMR and GT (Figure 50 and Table 92). The WKs for AI-assisted and GT and WKs for IMR and GT for F2&F3 vs other were equivalent, with WK for AI-assisted and GT being slightly higher than WK for IMR and GT (0.565 vs 0.546, respectively). For NAS $\geq$ 4 with $\geq$1 in each score category vs other the WK for AI-assisted and GT was significantly higher than WK for IMR and GT (0.632 vs 0.512, respectively, with a difference of 0.12 and 95% CI of (0.064, 0.166)). In addition, WKs for NASH resolution (defined as hepatocellular ballooning score of 0, lobular inflammation score of 0 or 1 and any steatosis score) between AI-assisted and GT vs WK for IMR and GT were computed. The WK for AI-assisted and GT was significantly higher than the WK for IMR and GT (0.532 vs 0.370, respectively, with a difference of 0.162 and 95% CI of (0.090, 0.209)).

Figure 50: WK comparisons for NASH aggregate component scores (F2&F3 vs other and NAS > 4 with >1 in each score category vs other) and NASH resolution.

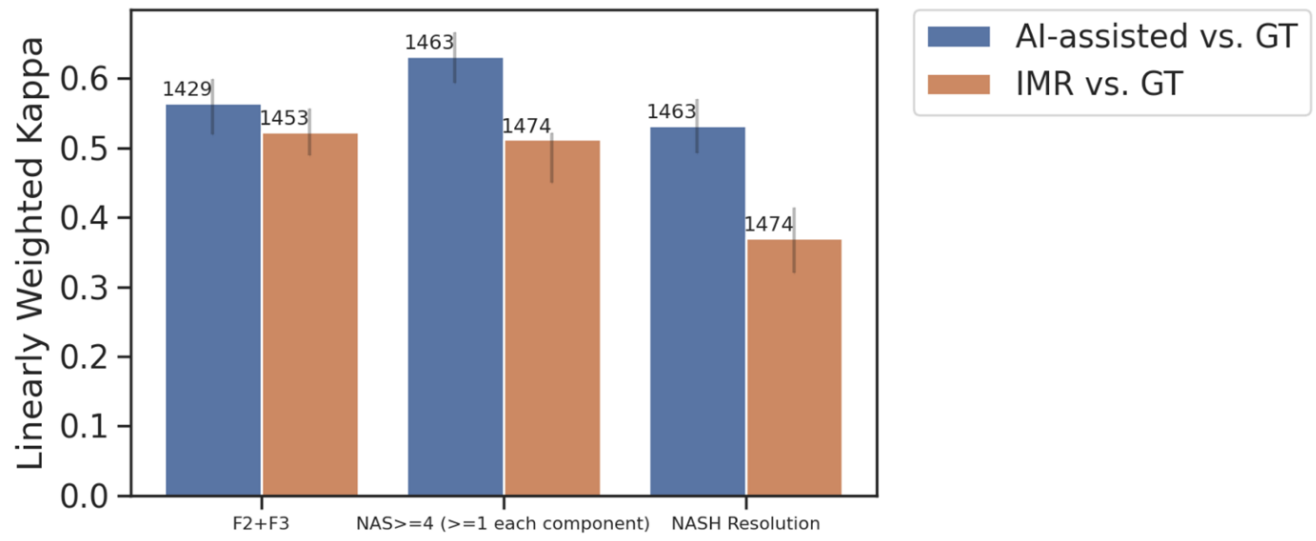


Table 92: WK comparisons for NASH aggregate component scores (F2&F3 vs other and NAS > 4 with >1 in each score category vs other) and NASH resolution

Feature	Modality	N	WK (95% CI)	Difference (95% CI)
F2&F3 vs other	AI-assisted vs GT	1429	0.565 (0.520, 0.599)	0.042 (-0.015, 0.093)
	IMR vs GT	1453	0.523 (0.485, 0.558)	
NAS $\geq$ 4 with $\geq$1 in each score category vs other	AI-assisted vs GT	1463	0.632 (0.593, 0.670)	0.12 (0.065, 0.169)
	IMR vs GT	1474	0.512 (0.478, 0.546)	
NASH resolution	AI-assisted vs GT	1463	0.532 (0.470, 0.542)	0.162 (0.090, 0.209)
	IMR vs GT	1474	0.370 (0.320, 0.414)	

- **Intra-reader agreement with and without AIM-NASH**

Since it was not always true that the same pathologists read a particular case with AIM-NASH and without (reads were obtained from a pool of pathologists), it was important to determine whether the overall results were similar for the subset of cases where the same pathologist reviewed the same slide for IMR and for AI-assisted. The number of cases where the same pathologist reviewed the same slide varied between 86 and 216 slides per component. Therefore, WKs were computed (Table 93) in this group. For all NASH components except for steatosis, the mean WK is higher for AI-assisted compared to IMR. For steatosis, the mean WK for AI-assisted and IMR is the same. In the study design, for the primary endpoint, the AI-assisted pathologist was not always part of the IMR for the same slide.

Table 93: Mean WKs for slides scored by the same pathologist for IMR and AI-assisted

Feature	Mean WK for AI-assisted vs GT	Mean WK for IMR vs GT
Steatosis	0.678	0.678
Lobular Inflammation	0.432	0.344
Hepatocellular Ballooning	0.585	0.475
Fibrosis	0.660	0.648

- **Inter-pathologist variability**

To assess the impact of the AIM-NASH algorithm on inter-pathologist variability, inter-rater agreement was computed between pathologists reviewing the same slide using AIM-NASH for each NASH component (

Table 94), using all the cases in the primary analysis as well as the spike-in cases. There were 79 slides with AI-assisted scores for NAS components (steatosis, lobular inflammation, and hepatocellular ballooning) and 74 slides with non-missing AI-assisted scores for fibrosis. Inter-rater agreement assessed using linearly WK, between pairs of pathologists reviewing the same slide ranged from 0.958 to 1 for steatosis, 0.973 to 1 for hepatocellular ballooning, and 0.906 to 1 for fibrosis. The inter-rater agreement for lobular inflammation was 1 for all pairwise WK. For the corresponding manual reads, pairwise agreement (for pathologists who read at least 10 of the shared cases), the WK ranged from 0.503 to 0.734 for steatosis, from 0.281 to 0.448 for hepatocellular ballooning, from 0.091 to 0.735 for fibrosis, and for lobular inflammation -0.047 to 0.466 (Table 95).

Table 94: Mean WKs for inter-reader agreement for AI-assisted

Feature	**Mean WK (range)**
Steatosis	0.986 (0.958-1)
Lobular inflammation	1
Hepatocellular ballooning	0.995 (0.973 - 1)
Fibrosis	0.958 (0.906 - 1)

Table 95: Mean WKs for inter-reader agreement for IMRs

Feature	**Mean WK (range)**
Steatosis	0.672 (0.503-0.734)
Lobular inflammation	0.229 (-0.047-0.466)
Hepatocellular ballooning	0.383 (0.281-0.448)
Fibrosis	0.493(0.091-0.735)

- **Disagreement with AIM-NASH**

In the AI-assisted workflow, pathologists were able to record if they disagreed with the AIM-NASH score by 1-stage or by 2 or more stages. However, to limit introduction of individual variability, pathologists could only change the final score if they disagreed by 2 or more stages for any component. The rates of 1-stage disagreement were as expected (

Table 96), given inter-reader variability demonstrated here and within the literature, but the 2-stage disagreements were quite low. Pathologists rejected the AIM NASH score with a 2-stage disagreement for each feature between 0.37%-1.83% of cases. (

Table 96).

Table 96: Percent of cases with 1 or 2-stage disagreement per NASH component

Feature	**1-stage Disagreement % (n/N)**	**2-stage Disagreement % (n/N)**
Steatosis	15.27 (246/1611)	0.37 (6/1611)
Lobular inflammation	16.63 (268/1611)	0.62 (10/1611)
Hepatocellular ballooning	22.10 (356/1611)	0.50 (8/1611)
Fibrosis	19.81 (314/1585)	1.83 (29/1585)

- **Accuracy of AIM-NASH Algorithm alone**

As stated in the proposed COU, the AIM-NASH algorithm aids pathologists in assessing NAS score and fibrosis stage in liver biopsies in NASH clinical trials. However, AIM-NASH algorithm only scores and IMR scores were also each evaluated for agreement to GT. For all 4 NASH components, non-inferiority was achieved (Figure 51 and

Table 97). For hepatocellular ballooning and lobular inflammation, superiority was also achieved. The difference in WK for AIM-NASH only and GT compared to WK for mean IMR and GT for hepatocellular ballooning was 0.148 (95% CI of (0.103, 0.193); NI $p<0.0001$) and for lobular inflammation was 0.119 (95% CI of (0.073, 0.166); NI $p<0.0001$) with a $p<0.0001$ for superiority for both components. The difference in WK for AIM-NASH only and GT compared to WK of mean IMR and GT for steatosis was 0.002 (95% CI of (-0.032, 0.037); NI $p<0.0001$) and for fibrosis was -0.009 (95% CI of (-0.044, 0.025); NI $p<0.0001$). Steatosis and fibrosis did not achieve superiority.

Figure 51: Accuracy analysis of AIM-NASH algorithm only in CV Population

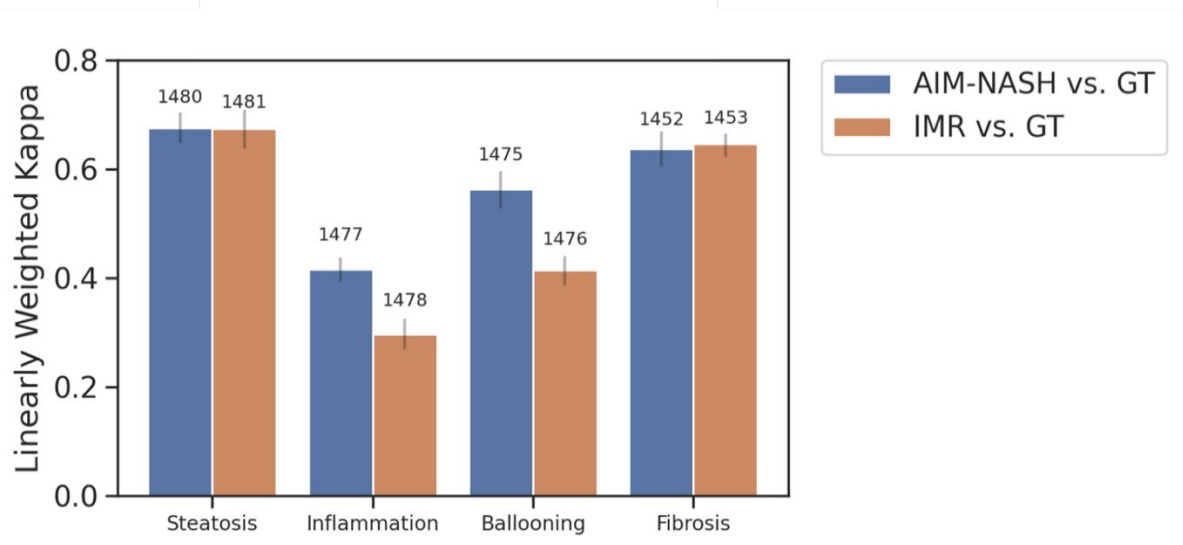


Table 97: Accuracy analysis for AIM-NASH algorithm only (w/out pathologist review)

Feature	Modality	N	WK (95% CI)	Difference (95% CI)	p-value for NI	p-value for superiority
Steatosis	AIM-NASH vs GT	1480	0.675 (0.649, 0.702)	0.002 (-0.032, 0.037)	<0.0001	0.444
	IMR vs GT	1481	0.674 (0.651, 0.694)			
Lobular inflammation	AIM-NASH vs GT	1477	0.416 (0.383, 0.450)	0.119 (0.073, 0.166)	<0.0001	<0.0001
	IMR vs GT	1478	0.297 (0.265, 0.329)			
Hepatocellular ballooning	AIM-NASH vs GT	1475	0.562 (0.526, 0.597)	0.148 (0.103, 0.193)	<0.0001	<0.0001
	IMR vs GT	1476	0.414 (0.385, 0.442)			
Fibrosis	AIM-NASH vs GT	1452	0.636 (0.608, 0.661)	-0.009 (-0.044, 0.025)	<0.0001	0.7045
	IMR vs GT	1453	0.645 (0.622, 0.665)			

4.7.12 Handling of Missing Data

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 GT, IMR, or AI-assisted were excluded from analysis of that NASH component.

4.7.13 Limitations

The slides used in this study were acquired from completed phase 2 and phase 3 clinical trials, with the addition of screen failures. The study was designed such that it is statistically powered with sample sizes able to observe differences by whole features, rather than by individual score levels and to provide evidence from multiple trials, demonstrating accuracy per the context of use population. The study is not powered at specific score levels within each histologic component, as this would have been a significantly higher sample size and not feasible to execute, and there are no published reference Kappas for manual pathology at each of these levels. However, secondary and exploratory analyses were performed to evaluate the success of the AI-assisted workflow at the individual score level compared to IMRs.

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 pathologists did not get to request a re-stain or a rescan of samples where they thought the sample was not of sufficient quality. This could have led to higher rates of samples

being deemed as inadequate, than in a prospective clinical trial setting, where these samples could be re-stained or rescanned.

4.7.14 Discussion and Conclusions

The CV dataset included biopsies from 3 separate NASH clinical trials with different drug candidates, 1 non-cirrhotic phase 2b cohort (Semaglutide/NCT04822181), 1 cirrhotic phase 2b cohort (Pegbelfermin/Falcon 2), and 1 non-cirrhotic phase 3 cohort (Obeticholic acid/Regenerate). This population is representative of current screened and enrolled NASH clinical trial populations. If the clinical trial population evolves (e.g. patients with earlier stage disease are screened), then we will monitor and determine if additional training is needed according to our change control processes. 2 out of the 3 clinical trials utilized in the CV dataset (Semaglutide/NCT04822181 and Obeticholic acid/Regenerate) recruited patients from the European countries, therefore making the validation data directly relevant to the European Union NASH patients. The biopsies were stained at various pathology labs per the original trial protocol and included baseline and follow-up timepoints from treated and placebo groups, as well as a screen failure subset. In total, over 1450 biopsies were evaluated by pathologists, either with the aid of the AIM-NASH scoring tool or manually (IMR), using the validated PathAI WSI viewer, with a minimum two-week wash-out between reads (for cases where the same pathologist read with both modality). These reads were compared to separate ground truth panels' scores for steatosis, lobular inflammation, hepatocellular ballooning, and fibrosis using the CRN scoring system (8), also performed on the validated viewer, in order to determine scoring accuracy in this large and diverse CV dataset with and without AIM-NASH (Figure 19).

Accuracy analysis demonstrated that AI-assisted scores are superior to manual pathologist scoring for lobular inflammation and hepatocellular ballooning and non-inferior for steatosis and fibrosis (Figure 43). Additionally, the WKs for all features for AI-assisted and ground truth were in the ranges of published CRN pathologists WKs (6,14). Accuracy per score within each NASH histologic component was also determined. For all grades of hepatocellular ballooning, AI-assisted pathologists showed superior accuracy over manual pathologists (Table 90). For steatosis AI-assisted reads had statistically higher WKs than manual pathologists for score 2 and 3. Similarly for lobular inflammation AI-assisted reads had statistically higher WKs than manual pathologists for score 1 and 3, indicating that AIM-NASH can improve inflammation scoring during trials considering these were the most commonly represented inflammation scores in the entire CV dataset (accounting for above 90% of biopsies per ground truth). For all other scores within histologic components, similar levels of accuracy with overlapping confidence intervals for AI-assisted and for manual pathologists were observed, with sometimes the mean being slightly higher for AI-assisted and for other score levels, manual pathologist mean WKs were slightly higher. It is worth noting that sample sizes for lobular inflammation 0 (n=12) and 3 (n=30) and fibrosis 0 (n=13) were low due to availability of data in scores less represented in NASH clinical trial populations, both screening and enrolled. Regardless, accuracy was similar to the average expert pathologist performance at these score levels, and granularity at the score level is not illustrated in the literature to our knowledge. The CV hepatocellular ballooning results are especially important, as identification and scoring of hepatocellular ballooning has shown to be a challenging and non-standardized task even among the CRN pathologists (14), AI-assisted results demonstrated here can greatly improve scoring accuracy during enrollment and follow-up timepoints in NASH clinical trials.

Secondary and exploratory analysis also demonstrated improved accuracy of AI-assisted scores compared to manual pathologist reads for specific NASH clinical trial relevant populations. Low stage fibrosis (F0&F1), fibrosis stage 2&3 (commonly selected for non-cirrhotic trials), and advanced fibrosis (F4,) are useful benchmarks when considering NASH disease improvement or progression or during enrollment for cirrhotic and non-cirrhotic

trials. AI-assisted reads demonstrated similar WKs to manual pathologist reads for all of those parameters. Similarly, the sum of the ordinal scores for steatosis, lobular inflammation, and hepatocellular ballooning (NAS) being greater than or equal to 4 (NAS≥4) is one of the main indicators for a probable NASH diagnosis, as well as commonly being a requirement for trial inclusion. Additionally, one component of the composite endpoint is NASH resolution, defined as a hepatocellular ballooning score of 0, lobular inflammation 0 or 1, and any score for steatosis. AI-assisted reads for NAS≥4 with ≥1 in each component category and for NASH resolution were statistically superior compared to unassisted manual pathologist's reads. This is an important indicator, in addition to the accuracy results per histologic component, that AIM-NASH can be a powerful tool in increasing and standardizing key aspects of trial scoring for enrollment and for FDA and EMA recommended endpoints.

Additional exploratory analyses for Falcon 2 and Regenerate trial subsets, where time point was available (baseline and post-baseline) described accuracy for each histologic component. For steatosis and fibrosis, AI-assisted pathologists performed similarly to unassisted manual pathologists with relatively high WKs. However, consistent with results thus far for hepatocellular ballooning and lobular inflammation, AI-assisted pathologists demonstrated significantly better accuracy at baseline and at post baseline compared to unassisted pathologists.

The clinical trial samples utilized in this study came from patients treated with Obeticholic Acid (Intercept), the FX receptor agonist which that works by decreasing lipid deposition in hepatocytes; Semaglutide (Novo Nordisk), the GLP1 receptor agonist which has been shown to reduces liver inflammation; and Pegbelfermin (BMS), the FGF21 analog which aims to regulate lipid and glucose metabolism in the liver. When each of the full trial subsets were analyzed, AI-assisted scores were more accurate for lobular inflammation and hepatocellular ballooning than unassisted pathologists for all 3 clinical trials (Falcon 2, Regenerate, and the Semaglutide ph2b). Accuracy was similar for steatosis and fibrosis for assisted and unassisted pathologists. Together, this data provides strong evidence that AIM-NASH can assist trial pathologists to result in better accuracy overall in scoring biopsies for drug candidates with a variety of mechanisms of action.

Inter-reader variability is also a significant challenge in NASH trials, often resulting in inconsistent scoring if more than one pathologist provides reads for different subjects or timepoints. Additionally, when multiple readers are used per case to determine consensus scores, this large inter-rater variability often causes a delay in screening patients due to the need for many consensus sessions needed to resolve discrepancies. Additionally, in these consensus sessions, there may be a dominant voice which can bias results to one pathologist's scores. Finally, the high inter-rater variability makes it difficult to compare across trials or reproduce results from phase 2 to phase 3 if a different pathologist or consensus approach is used. For a subset of cases, multiple pathologists independently used AI-assisted scoring and the agreement between assisted pathologists was compared to that between unassisted pathologists for each histologic component. Strikingly, for all NASH histologic components (considering any score change of 1 or more as a disagreement), the inter-reader WKs were all greater than 0.9 (

Table 94), which are significantly higher than the WKs for unassisted manual reads for the same slides (which ranged from 0.229 to 0.672; Table 95) and published inter-reader WKs for NASH components (ranging from 0.328 to 0.77 (6,14). Additionally, in cases where the same pathologist read with and without AI-assistance, the assisted workflow brought the pathologist closer to the ground truth for all histologic components except steatosis (Table 93). For steatosis, the AI-assisted reads were equivalent to the unassisted reads. These results indicate that using AI-assisted workflows in clinical trials could decrease the time and cost of trials as only one pathologist is needed to achieve highly accurate and consistent reads, instead of the current gold standard of 3 pathologists which can still be subject to inconsistent results for different groups of pathologists and due to temporal bias during enrollment vs. follow-up evaluations.

The AI-assisted workflow consists of the pathologist review of the sample, stain, and scan adequacy, as well as review of AIM-NASH generated scores. To reduce the addition of scoring variability during pathologist review, the pathologist is only able to reject the AIM-NASH score if they disagree with the score by 2 or more for any histologic component. Considering the low rate of algorithm score rejection by the primary pathologist (approximately 1.8%), additional accuracy analysis was performed for AIM-NASH scores only (without the pathologist review) for all clinical trial samples utilized in this CV study. These algorithm-only accuracy data can strongly support the overall analytical performance evidence package for AIM-NASH, since the dataset includes a very large sample size with a wide range of scores and disease activity, multiple drug candidates with various modes of actions, and variability in pre-analytical factors, such as staining. Strikingly, the results are very similar to AI-assisted results. AIM-NASH results alone are superior to unassisted pathologist reads for lobular inflammation and hepatocellular ballooning, and non-inferior for steatosis and fibrosis. Interestingly, the WKs for AIM-NASH only vs ground truth and AI-assisted vs ground truth for all NASH components are very similar (steatosis 0.675 vs 0.677, lobular inflammation 0.416 vs 0.419, hepatocellular ballooning 0.562 vs 0.563 and fibrosis 0.636 vs 0.654, respectively).

These findings demonstrate that utilizing AIM-NASH with pathologist review can be a powerful tool in NASH clinical trials, capable of accurately and consistently evaluating patients for trial inclusion criteria and in measuring the success or failure to meet study endpoints, which is currently an unmet need.

The body of evidence generated in this clinical validation study indicates that utilizing AIM-NASH to assist pathologists in NASH scoring could significantly help to solve the current challenges in reducing reader variability and allow for consistent and accurate histologic scoring for patient enrollment and determination of histologic-based primary endpoints currently recommended by FDA and EMA and used for accelerated approval. This is especially important in considering the substantial challenges in the standardized identification and scoring of hepatocellular ballooning and lobular inflammation (6,14,26), and the fact that hepatocellular ballooning and lobular inflammation scores are key in evaluating both disease activity and resolution of NASH, AI-assisted results demonstrated here can greatly improve pathologist scoring in NASH trials, and, therefore, help to better identify effective drug treatments for NASH patients.

4.8 Qualification of Pathologists for Validation studies

All PathAI network pathologists are qualified board-certified pathologists that have undergone study-specific training on the viewing platform(s) and have demonstrated proficiency in NASH scoring prior to participating in any validation studies Pathologists (Table 98) are required to meet the following selection criteria:

- Board certification in pathology as evidenced by documentation of The American Board of Medical Specialties (ABMS) certification or equivalent certification in the country for which they are practicing.
- Liver pathology subspecialty as evidenced by liver pathology fellowship training and/or significant ongoing clinical experience.

Table 98: PathAI Network Pathologists Qualifications

Pathologist	Fellowship	Years of Experience	# of NASH cases/month	Study
Pathologist 1	GI and Surgical Pathology	4	50	OV evaluating
Pathologist 2	GI Pathology	20	4	OV evaluating/PV GT
Pathologist 3	GI and Liver Pathology	10	5	OV evaluating/AV, CV GT/ PV GT
Pathologist 4	Surgical Pathology	>30	30-40	OV enrollment/AV, CV GT
Pathologist 5	GI and Liver Pathology	3	20	AV, CV IMR and AI-assisted
Pathologist 6	Surgical Pathology	15	5-10	AV, CV IMR and AI-assisted
Pathologist 7	GI and Hepatobiliary Pathology	7	50	AV, CV IMR and AI-assisted
Pathologist 8	GI Pathology	>20	10	AV, CV IMR
Pathologist 9	GI and Liver Pathology	18	12	AV, CV IMR
Pathologist 10	GI Pathology	4	4	AV, CV IMR and AI-assisted
Pathologist 11	GI and Liver Pathology	12	30	AV, CV IMR and AI-assisted
Pathologist 12	GI Pathology	15	10	AV, CV IMR and AI-assisted
Pathologist 13	GI Pathology	20	10	AV, CV GT
Pathologist 14	GI Pathology	11	15-20	AV, CV GT/ PV GT
Pathologist 15	GI and Liver Pathology	2	2	AV, CV GT
Pathologist 16	Surgical Pathology	5	5-10	PV
Pathologist 17	Surgical Pathology	20	10-15	PV
Pathologist 18	GI and Liver Pathology	2	30	PV
Pathologist 19	GI Pathology	2	2	PV
Pathologist 20	GI Pathology	2	10	PV
Pathologist 21	GI Pathology	15	20	PV

AI-assisted - AIM-NASH review, AV - analytical validation, CV - clinical validation, PV - platform validation for AISight Clinical Trials and AISight Translational WSI viewers, GI - gastrointestinal, GT - ground truth, IMR - individual manual reads, OV - overlay validation

Proficiency testing was performed using 50 H&E slides and 50 trichrome slides from a completed Phase 2B study and required pathologists to provide a score for each NASH component (steatosis, lobular inflammation, hepatocellular ballooning, and CRN fibrosis). NASH component scores by the central pathologist for the phase 2B study were also available. The linearly weighted Kappa statistic for each pathologist score compared to the consensus median score and the consensus central pathologist median score were calculated. No information from the original Phase 2B study was available to the pathologists (including time point or treatment arm). Pathologists were chosen to participate in validation studies based on their experience, proficiency testing results and availability.

Consensus reads: the gold standard in this trial is reference reads from three trained and qualified pathologists who prospectively read a balanced set of slides with a wide range of quality. If 2 out of 3 pathologists agree,

this will be the final score. If all 3 disagree, a consensus call is scheduled. The analytical and clinical validation reads are compared against the gold standard and not validated against clinical outcome. The full Pathologist Qualification and Proficiency protocol is available in Appendix IIa.

Evidence from Published Literature

CASE STUDY 1:

Lead Author/Title: Diane Shevell, Comparison of manual vs machine learning approaches to liver biopsy scoring for NASH and fibrosis: a post hoc analysis of the FALCON 1 study.

Conference: Poster presentation at the American Association for the Study of Liver Diseases Meeting, 2021

Clinical trial (Sponsor): NCT03486899 (Bristol Myers Squibb)

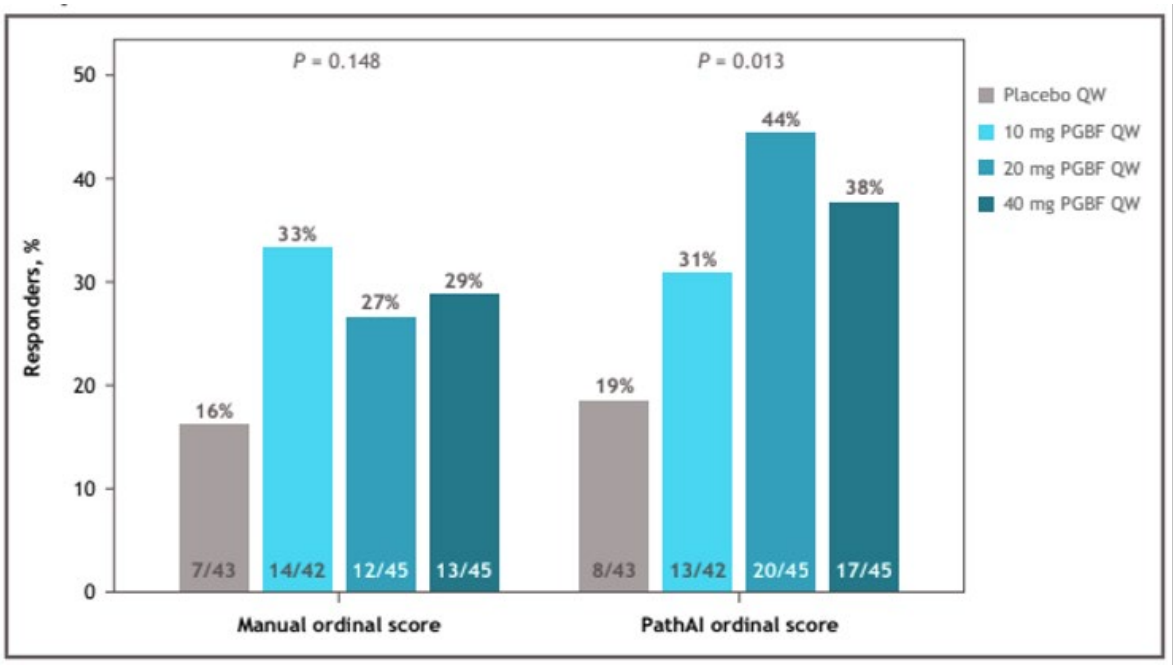
Method: Whole slide images of H&E- and Masson's trichrome-stained liver biopsy tissue from subjects enrolled in the FALCON 1 NASH clinical trial (Ph2b study of pegbelfermin for treatment of NASH in adults with Stage 3 Liver Fibrosis) were read by the study's Central Pathologist and by AIM-NASH. Response rates per treatment group were recorded and compared across the histologic evaluation methodologies. Primary endpoint responders were patients with ≥1-stage improvement in NASH CRN fibrosis stage without NASH worsening, or NASH improvement with no worsening of fibrosis, at Week 24.

Key results:

- AIM-NASH revealed a statistically significant difference in the proportion of primary endpoint responders in Treatment vs. Placebo groups (
- Figure 52).
- Central Pathologist scoring did NOT reveal a statistically significant difference in the proportion of primary endpoint responders in Treatment vs. Placebo groups (
- Figure 52).
- AIM-NASH revealed a statistically significant difference in the proportion of patients who demonstrated ≥1-grade reduction in lobular inflammation in Treatment vs. Placebo groups.
- Central Pathologist scoring did NOT reveal a statistically significant difference in the proportion of patients who demonstrated ≥1-grade reduction in lobular inflammation in Treatment vs. Placebo groups.
- AIM-NASH revealed a statistically significant difference in the proportion of patients who demonstrated ≥1-grade reduction in hepatocellular ballooning in Treatment vs. Placebo groups.
- Central Pathologist scoring did NOT reveal a statistically significant difference in the proportion of patients who demonstrated ≥1-grade reduction in hepatocellular ballooning lobular inflammation in Treatment vs. Placebo groups.

Figure 52: AIM-NASH vs. Central Pathologist detection of primary endpoint response in a Ph2 study of pegbelfermin for treatment of NASH with CRN Fibrosis Stage 3. Primary endpoint responders were patients with ≥ 1 stage NASH CRN fibrosis improvement without NASH worsening or NASH improvement

with no worsening of fibrosis at week 24. Cochran-Armitage test for trend was used to compare PGBF vs placebo. NASH, nonalcoholic steatohepatitis; PGBF, pegbelfermin; QW, once weekly.



CASE STUDY 2:

Lead Author/Title: Stephen Harrison 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 nonalcoholic steatohepatitis.

Conference: Poster presentation at European Association for the Study of Liver Diseases Meeting, London, England, 2022.

Clinical Trial (Sponsor): NCT02912260 (Madrigal Pharmaceuticals)

Method: Whole slide images of H&E- and Masson's trichrome-stained liver biopsy tissue from subjects enrolled in the Ph2 study of MGL-3196 (resmetirom) for treatment of NASH in patients with NASH CRN Fibrosis stages 1-3 were read by the study's central pathologist, a second expert NASH pathologist, and AIM-NASH. Response rates per treatment group were recorded and compared across the histologic evaluation methodologies. Endpoints evaluated for comparison between the methodologies included the proportion of patients in resmetirom vs. Placebo groups who demonstrated 1) >= 2-point reduction in NAFLD Activity Score (NAS), and 2) NASH resolution without worsening of fibrosis.

Key results:

AIM-NASH detected a statistically significant difference between response rates in the resmetirom vs. placebo subject groups **(**Table 99**)**.

Both the Central Reader and the independent expert NASH pathologist ("Reader 2") also detected statistically significant differences between response rates in the resmetirom vs. placebo subject groups **(**Table 99**)**.

Table 99: AIM-NASH vs. Pathologist detection of endpoint response in Ph2 study of MGL-3196 for treatment of NASH with CRN Fibrosis Stages 1-3. Consistent with the Central Pathologist and Reader 2, AIM-NASH detected a significantly greater treatment response in the resmetirom-treated group relative to placebo.

Endpoint	Scorer	Resmetirom response rate	Placebo response rate	p-value
≥2-point improvement in NAS	AIM-NASH	0.41	0.19	0.0327
	Central reader	0.56	0.26	0.0044
	Reader 2	0.42	0.19	0.0321
NASH resolution without worsening of fibrosis	AIM-NASH	0.26	0.07	0.0301
	Central reader	0.25	0.06	0.0226
	Reader 2	0.21	0.03	0.0190

CASE STUDY 3:

Lead Author/Title: Stephen Harrison, Artificial intelligence-powered digital pathology model supports that fibrosis is reduced by semaglutide in patients with NASH.

Conference: American Association for the Study of Liver Diseases Meeting, 2021

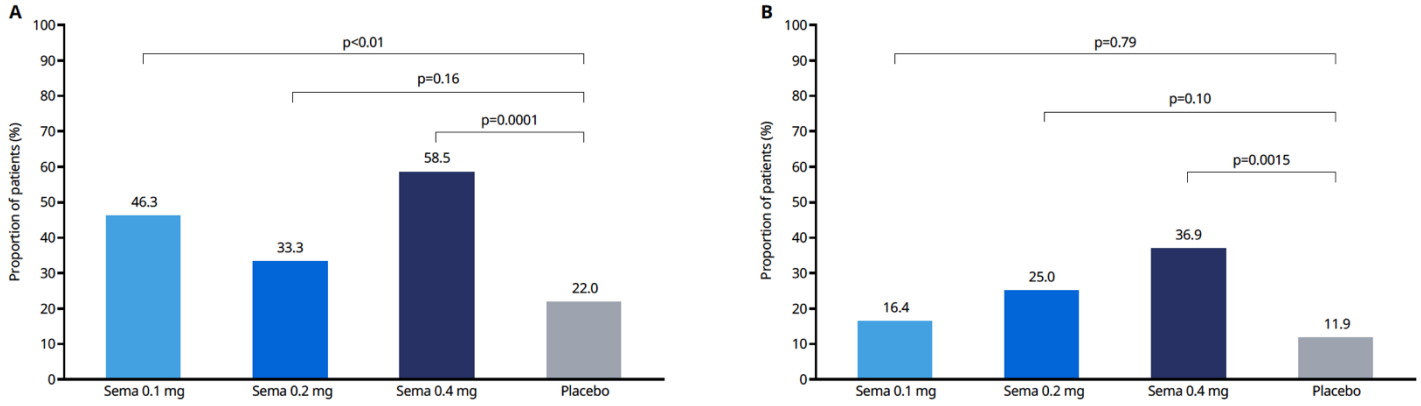
Clinical Trial: NCT02970942 (Novo Nordisk)

Method: Whole slide images of H&E- and Masson's trichrome-stained liver biopsy tissue from subjects enrolled in the Ph2 study of semaglutide for treatment of NASH in patients with NASH CRN Fibrosis stages 1-3 were read by the study's Central Pathologists (N=2) and AIM-NASH. Response rates per treatment group were recorded and compared across the histologic evaluation methodologies. The endpoint evaluated was the proportion of patients with NASH resolution without fibrosis worsening. **Key results:**

AIM-NASH detected a dose-related response in subjects treated with semaglutide, where increasing dosages of semaglutide resulted in increasingly improved drug response **(Figure 53, Panel B).**

Central Pathologist scoring did NOT detect a dose-related drug response in subjects treated with semaglutide **(Figure 53, Panel A)**.

Figure 53: Dose-related drug response detected via Central Pathologists vs. AIM-NASH in Ph2 study of semaglutide for treatment of NASH with CRN Fibrosis Stages 1-3. (A) Dose-related drug response is not detected by Central Pathologist scoring. (B) Dose-related dr



CASE STUDY 4:

Lead Author, Title: Rohit Loomba, Comparison of the effects of semaglutide on liver histology in patients with non-alcoholic steatohepatitis cirrhosis between machine learning model assessment and pathologist evaluation.

Conference: Poster presentation at the American Association for the Study of Liver Diseases Meeting, Washington, DC, 2022.

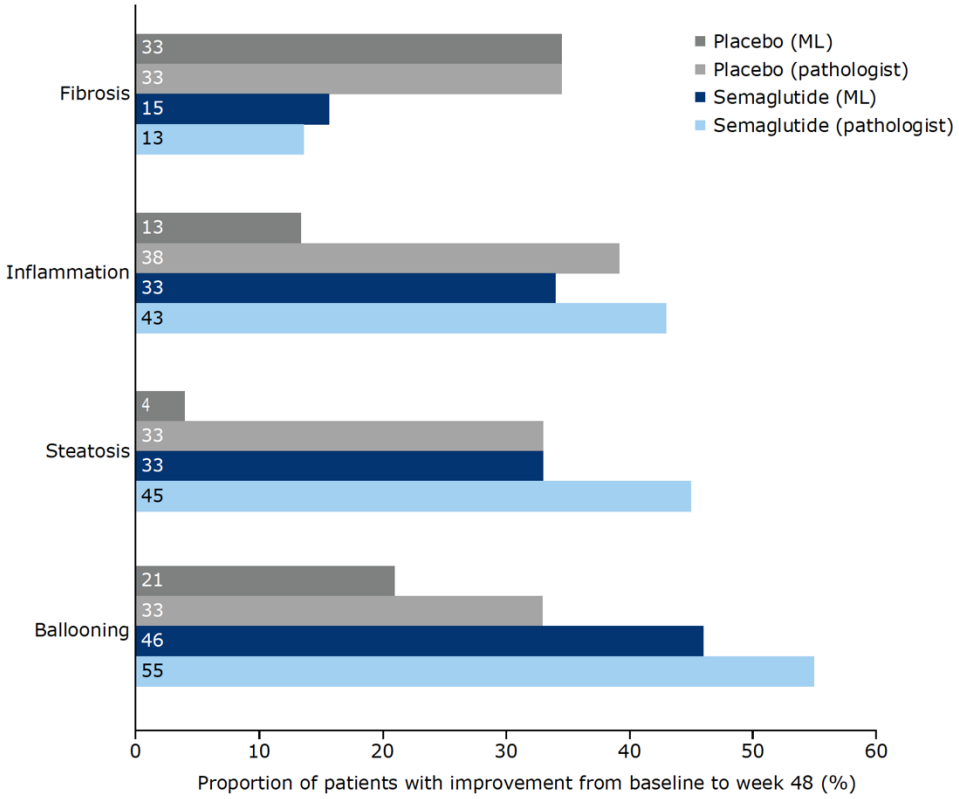
Clinical Trial (Sponsor): NCT03987451 (Novo Nordisk)

Method: Whole slide images of H&E- and Masson's trichrome-stained liver biopsy tissue from subjects enrolled in the Ph2 study of semaglutide for treatment of NASH in patients with cirrhosis were read by the study's Central Pathologist and AIM-NASH. The proportion of patients across treatment groups who showed improvement in steatosis, ballooning, lobular inflammation, and fibrosis was recorded and compared across the histologic evaluation methodologies.

Key result:

AIM-NASH detected significantly lower placebo response than the Central Pathologist for steatosis (4% vs. 33%), ballooning (21% vs. 33%), and lobular inflammation (13% vs. 38%) **(**Figure 54**)**.

Figure 54: Histologic feature-specific response rates across Treated vs. Placebo subjects, as measured by the Central Pathologist vs. AIM-NASH, in a Ph2 study of semaglutide for treatment of NASH with cirrhosis. For Inflammation, Steatosis, and Ballooning, AIM-NA



4.9 Remaining Gaps

In the event that the context of use for the AIM-NASH algorithm is expanded beyond the use of only the Aperio AT2 scanner, a planned bridging study will be designed to address the gap that exists due to the AIM-NASH algorithm having been trained and validated on just one scanner at a single magnification. The objectives of this study are to assess AIM-NASH model performance and reproducibility between commonly used whole slide scanners. In this study, PathAI will assess the AIM-NASH scoring agreement between two magnifications (20x and 40x) at the case (one H&E and one trichrome slide) level by assessing the outputs (NAS score and CRN Fibrosis stage) of WSI generated at different magnifications. These same scoring agreements will be assessed between the Aperio AT2, Aperio GT450, Philips Ultra Fast Scanner, and Ventana DP200.

4.10 Conclusions

The variability, both inter- and intra-reader, demonstrated by expert NASH pathologists, both in the literature and, furthermore, in NASH clinical trials, has continued to persist, likely for multiple reasons. The CRN scoring system is accepted by both EMA and FDA for use in enrollment and in calculating composite endpoints for accelerated or conditional approval. Variability in expert pathologist reads likely stems from a combination of

factors in two areas: first in interpretation of individual features, and second, in estimating the quantity of these features according to the scoring systems over an entire case. Feature interpretation is variable because there is high subjectivity and a lack of standardization in the definition for identification and quantification of certain key histologic components, notably hepatocellular ballooning, foci of lobular inflammation, and fibrosis staging (especially for borderline cases). Unclear guidelines for quantification for different components (e.g., none vs few vs many) compound this problem. Integrating these features from hundreds of high-powered fields that comprise a case likely further increases the variability. The variability between pathologists has not improved over time and Kappas can differ widely, depending on the number of pathologists being compared (e.g., between a pair of pathologists, or average Kappa over a larger number of experts (6,8,14), and the composition of pathologist panels (e.g., two pathologists with a more conservative definition of ballooning or a more balanced panel who will fall somewhere in the middle after consensus discussion).

This scoring system, developed as aid for diagnosing NASH in the clinic, plays a different role in ph2 and ph3 NASH clinical trials, one that depends more on the precision of scoring over time for enrollment and for endpoint evaluation. This can be seen by the variable placebo rates demonstrated across phases and across different drug candidate trials with similar inclusion criteria, and variability in scoring during re-reads (when enrollment biopsies are read again to blind pathologists for follow-up biopsy assessments). Additionally, some trials have used two readers during enrollment, and screen failures have changed depending on the reader being utilized, or consensus method incorporated. This inter-reader variability is also demonstrated when studies utilizing the 2+1 consensus approach (2 pathologists, if disagreeing on any component score, go to panel consensus call and only use an arbitrator if they can't come to an agreement) experience up to 85% initial disagreement for at least one component. Finally, there remain inconsistencies in using a consensus panel, depending on the composition of the panels (2 conservative pathologists vs. 2 more generous with their ballooning definition), the temporal bias in scoring during enrollment vs. follow-up timepoints, and the consensus workflow (2+1, or 3 pathologists where 2 out of 3 agreement is used plus consensus calls for completely discrepant cases, etc.), as well as other factors (23). There's been research to show that, due to the inherent variability in scoring, a well powered study (>90%) may see drastic reductions in power to as low as 40%, inhibiting the ability to detect change (14).This is key in ph2b trials which usually have too few sample sizes as is and could potentially result in promising drug candidates never advancing to ph3 or not receiving approval overall. Additionally, using multiple pathologists is very costly and delays enrollment substantially, potentially at the risk of not including some patients who could have benefited.

The above unmet need is urgent, with the prevalence of NAFLD exceeding 25% in European adults, and NASH estimated at about 3% and rising (35,36). There are no currently available therapies, and no precise standard with which other diagnostic tools (e.g., histologic or non-invasive tests) can be compared or validated. AI-based tools have the potential to solve many of the issues around precision, including standardized, accurate and reproducible scoring, within and across trials. Multiple pathologists can assess biopsies on validated whole slide imaging viewers, both for sample adequacy/evaluability and for overall diagnosis and additional findings, while using AI tools to efficiently provide the associated accurate and consistent scores needed. In this body of data, the AISight Clinical Trials platform has been validated to be equivalent to glass for NASH trial reads. The AIM-NASH outputs have been validated according to their proposed use with highly representative trial datasets, both variable in disease activity, stain, scanning site, and drug candidate. Specifically, the overlays presented to the pathologist, identifying key areas that the model is predicting as artifact, steatosis, ballooning, inflammation, and fibrosis, have been validated by multiple pathologist readers on a frames level, demonstrating 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 histologic component score outputs for AIM-NASH have been shown to be as

accurate, or more, than the expert pathologist readers, compared to ground truth, in the variable AV and CV datasets which included samples from four trials. These trials included varying staining, inclusion criteria, time-points, and drug candidates with different mechanisms of action. In the external repeatability and reproducibility studies, AIM-NASH was consistently superior to study inter-pathologist agreements, as well as literature inter- and intra-pathologist agreements. During CV, pathologists used AIM-NASH to score representative biopsies across three trials. Consistently, for the histologic components historically most difficult to score (ballooning and inflammation), AIM-NASH brought individual pathologists closer to ground truth reads while maintaining high levels of accuracy for steatosis and fibrosis. Strikingly, even considering 1-point disagreements, AIM-NASH substantially decreased inter-pathologist variability compared to unassisted reads, with all component Kappas ranging from 0.9 to 1 for AIM-NASH assisted reads. Together, the above data strongly supports the use of AIM-NASH by pathologists in trials and can play a strong role in solving the accuracy and precision gaps in NASH assessment, while guiding pathologists for an efficient evaluation to result in a standardized and reproducible score within and across trials. This in turn could significantly benefit NASH patients in helping to bring truly effective therapies to market.

5 Appendices

5.1 Appendix I. References

1. Younossi ZM, Koenig AB, Abdelatif D, Fazel Y, Henry L, Wymer M. Global epidemiology of nonalcoholic fatty liver disease—Meta-analytic assessment of prevalence, incidence, and outcomes. Hepatology. 2016;64(1).

2. Friedman SL, Neuschwander-Tetri BA, Rinella M, Sanyal AJ. Mechanisms of NAFLD development and therapeutic strategies. Vol. 24, Nature Medicine. 2018.

3. Brunt EM, Janney CG, Di Bisceglie AM, Neuschwander-Tetri BA, Bacon BR. Nonalcoholic steatohepatitis: A proposal for grading and staging the histological lesions. American Journal of Gastroenterology. 1999;94(9).

4. FDA-NIH Biomarker Working Group. BEST (Biomarkers , EndpointS , and other Tools) Resource [Internet]. Updated, September 25. 2016.

5. Tong X fei, Wang Q yi, Zhao X yan, Sun Y meng, Wu X ning, Yang L ling, et al. Histological assessment based on liver biopsy: the value and challenges in NASH drug development. Vol. 43, Acta Pharmacologica Sinica. 2022.

6. Brunt EM, Kleiner DE, Wilson LA, Sanyal AJ, Neuschwander-Tetri BA. 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. 2019;70(2).

7. Angulo P, Kleiner DE, Dam-Larsen S, Adams LA, Bjornsson ES, Charatcharoenwitthaya P, et al. Liver fibrosis, but no other histologic features, is associated with long-term outcomes of patients with nonalcoholic fatty liver disease. Gastroenterology. 2015;149(2).

8. Kleiner DE, Brunt EM, Van Natta M, Behling C, Contos MJ, Cummings OW, et al. Design and validation of a histological scoring system for nonalcoholic fatty liver disease. Hepatology. 2005;41(6).

9. Filozof C, Chow SC, Dimick-Santos L, Chen YF, Williams RN, Goldstein BJ, et al. Clinical endpoints and adaptive clinical trials in precirrhotic nonalcoholic steatohepatitis: Facilitating development approaches for an emerging epidemic. Hepatol Commun. 2017;1(7).

10. Sanyal AJ, Brunt EM, Kleiner DE, Kowdley K V., Chalasani N, Lavine JE, et al. Endpoints and clinical trial design for nonalcoholic steatohepatitis. In: Hepatology. 2011.

11. FDA. Nonalcoholic Steatohepatitis with Compensated Cirrhosis: Developing Drugs for Treatment Guidance for Industry - Draft. 2019.

12. FDA. Noncirrhotic Nonalcoholic Steatohepatitis With Liver Fibrosis: Developing Drugs for Treatment Guidance for Industry. 2018.

13. EMA. Reflection paper on regulatory requirements for the development of medicinal products for chronic non-infectious liver diseases (PBC, PSC, NASH) (EMA/CHMP/299976/2018). European Medicines Agency. 2018;44(November 2018).

14. Davison BA, Harrison SA, Cotter G, Alkhouri N, Sanyal A, Edwards C, et al. Suboptimal reliability of liver biopsy evaluation has implications for randomized clinical trials. J Hepatol. 2020;73(6).

15. Merriman RB, Ferrell LD, Patti MG, Weston SR, Pabst MS, Aouizerat BE, et al. Correlation of paired liver biopsies in morbidly obese patients with suspected nonalcoholic fatty liver disease. Hepatology. 2006;44(4).

16. Juluri R, Vuppalanchi R, Olson J, Ünalp A, Van Natta ML, Cummings OW, et al. Generalizability of the nonalcoholic steatohepatitis clinical research network histologic scoring system for nonalcoholic fatty liver disease. J Clin Gastroenterol. 2011;45(1).

17. Pavlides M, Birks J, Fryer E, Delaney D, Sarania N, Banerjee R, et al. Interobserver variability in histologic evaluation of liver fibrosis using categorical and quantitative scores. Am J Clin Pathol. 2017;147(4).

18. Harrison SA, Alkhouri N, Davison BA, Sanyal A, Edwards C, Colca JR, et al. Insulin sensitizer MSDC-0602K in non-alcoholic steatohepatitis: A randomized, double-blind, placebo-controlled phase IIb study. J Hepatol. 2020;72(4).

19. Muehlematter UJ, Daniore P, Vokinger KN. Approval of artificial intelligence and machine learning-based medical devices in the USA and Europe (2015–20): a comparative analysis. Vol. 3, The Lancet Digital Health. 2021.

20. Chalasani N, Younossi Z, Lavine JE, Charlton M, Cusi K, Rinella M, et al. The diagnosis and management of nonalcoholic fatty liver disease: Practice guidance from the American Association for the Study of Liver Diseases. Hepatology. 2018;67(1).

21. Loomba R, Sanyal AJ. The global NAFLD epidemic. Vol. 10, Nature Reviews Gastroenterology and Hepatology. 2013.

22. Kleiner DE, Brunt EM, Wilson LA, Behling C, Guy C, Contos M, et al. Association of Histologic Disease Activity With Progression of Nonalcoholic Fatty Liver Disease. JAMA Netw Open. 2019;2(10).

23. Sanyal A, Loomba R, Anstee Q, Ratziu V, Shah A, Natha M, et al. Minimizing Variability and Increasing Concordance for NASH Histological Scoring in NASH Clinical Trials. In AASLD; 2021.

24. Newsome PN, Buchholtz K, Cusi K, Linder M, Okanoue T, Ratziu V, et al. A Placebo-Controlled Trial of Subcutaneous Semaglutide in Nonalcoholic Steatohepatitis. New England Journal of Medicine. 2021;384(12).

25. Gawrieh S, Knoedler DM, Saeian K, Wallace JR, Komorowski RA. Effects of interventions on intra- and interobserver agreement on interpretation of nonalcoholic fatty liver disease histology. Ann Diagn Pathol. 2011;15(1).

26. Brunt EM, Clouston AD, Goodman Z, Guy C, Kleiner DE, Lackner C, et al. Complexity of ballooned hepatocyte feature recognition: Defining a training atlas for artificial intelligence-based imaging in NAFLD. J Hepatol. 2022;76(5).

27. Brunt EM. Liver biopsy reliability in clinical trials: Thoughts from a liver pathologist. Vol. 73, Journal of Hepatology. 2020.

28. Sanyal AJ, Friedman SL, Mccullough AJ, Dimick-Santos L. Challenges and opportunities in drug and biomarker development for nonalcoholic steatohepatitis: Findings and recommendations from an American Association for the Study of Liver Diseases-U.S. Food and Drug Administration Joint Workshop. Hepatology. 2015;61(4).

29. Evans AJ, Brown RW, Bui MM, Chlipala EA, Lacchetti C, Milner DA, et al. Validating Whole Slide Imaging Systems for Diagnostic Purposes in Pathology: Guideline Update From the College of American Pathologists in Collaboration With the American Society for Clinical Pathology and the Association for Pathology Informatics. Arch Pathol Lab Med. 2021;

30. Krizhevsky, A., Sutskever, I. H. "Imagenet classification with deep convolutional neural network", in Advances in Neural Information Processing Systems, p. 1097-1105. Elsevier Ltd. 2012;

31. Sagawa S, Koh PW, Hashimoto TB, Liang P. Distributionally Robust Neural Networks for Group Shifts: On the Importance of Regularization for Worst-Case Generalization. 2019 Nov 20;

32. Heinze-Deml C, Meinshausen N. Conditional variance penalties and domain shift robustness. Mach Learn. 2021;110(2).

33. Borowsky AD, Glassy EF, Wallace WD, Kallichanda NS, Behling CA, Miller D V., et al. Digital whole slide imaging compared with light microscopy for primary diagnosis in surgical pathology: A multicenter, double-blinded, randomized study of 2045 cases. Arch Pathol Lab Med. 2020;144(10).

34. Mukhopadhyay S, Feldman MD, Abels E, Ashfaq R, Beltaifa S, Cacciabeve NG, et al. Whole Slide Imaging Versus Microscopy for Primary Diagnosis in Surgical Pathology: A Multicenter Blinded Randomized Noninferiority Study of 1992 Cases (Pivotal Study). American Journal of Surgical Pathology. 2018;42(1).

35. Tacke F. Non-alcoholic steatohepatitis (NASH): Definition, natural history and current therapeutic interventions. EMA Workshop on Liver Diseases . London; 2018.

36. Cholongitas E, Pavlopoulou I, Papatheodoridi M, Markakis GE, Bouras E, Haidich AB, et al. Epidemiology of nonalcoholic fatty liver disease in europe: A systematic review and meta-analysis. Ann Gastroenterol. 2021;34(3).

5.2 Appendix II: Pathologist Qualification and Proficiency Protocol

5.2.1 Appendix IIa: Analytical and Clinical Validation Pathologist Qualification and Proficiency

5.2.2 Appendix IIb: AISight CTS and Translational Pathologist Proficiency

5.2.3 Appendix IIc: Overlay Validation Pathologist Proficiency

5.3 Appendix III: Analytical Validation

5.3.1 Appendix IIIa: Protocol

5.3.2 Appendix IIIb: Report

5.3.3 Appendix IIIc: Example CRF's

5.4 Appendix IV: Clinical Validation

5.4.1 Appendix IVa: Protocol

5.4.2 Appendix IVb: Report

5.4.3 Appendix IVc: Example CRF's

5.5 Appendix V: AISight Clinical Trial Platform Validation

5.5.1 Appendix Va: Protocol

5.5.2 Appendix Vb: Report

5.5.3 Appendix Vc: Example CRF's

5.6 Appendix VI: AISight Translational (Slides) Platform Validation

5.6.1 Appendix VIa: Protocol

5.6.2 Appendix VIb: Report

5.6.3 Appendix VIc: Example CRF's

5.7 Appendix VII: Overlay Validation

5.7.1 Appendix VIIa: Protocol

5.7.2 Appendix VIIb: Report

5.7.3 Appendix VIIc: Example CRF's

5.8 Appendix VIII: AIM-NASH Model Revision History Table

5.9 Appendix IX: Integrated Analytical Verification

5.9.1 Appendix IXa: Protocol

5.9.2 Appendix IXb: Report

5.10 Appendix X: Standalone Validation

5.10.1 Appendix Xa: Protocol

5.10.2 Appendix Xb: Report