EUROPEAN MEDICINES AGENCY
SCIENCE MEDICINES HEALTH

24 April 2025
EMADOC-1700519818-1761332 Corr.1
Committee for Medicinal Products for Human Use (CHMP)

Qualification opinion for Artificial Intelligence-Based Measurement of Non-alcoholic Steatohepatitis Histology in Liver Biopsies to Determine Disease Activity in NASH/MASH Clinical Trials

Draft agreed by Scientific Advice Working Party (SAWP)	31 October 2024
Adopted by CHMP for release for consultation	14 November 2024[1]
Start of public consultation	6 December 2024[2]
End of consultation (deadline for comments)	24 January 2025
Adoption by CHMP	27 February 2025

Keywords	NASH/MASH clinical trials, Artificial Intelligence based tool to aid trial pathologists in scoring liver biopsies, AIM-NASH, patient inclusion, evaluation of study endpoints, NAS, fibrosis stage, Qualification of Novel Methodology

[1] Last day of relevant committee meeting
[2] Date of publication on the EMA public website

Qualification Opinion agreed by CHMP

Based on the evidence presented, CHMP considers that the proposed Artificial Intelligence-Based Measurement of Non-alcoholic Steatohepatitis Histology (AIM-NASH) tool can be used in metabolic dysfunction associated steatohepatitis (MASH) clinical trials as an aid to trial pathologists in scoring the liver biopsies at enrolment and follow-up visits.

Agreed Context of Use (CoU):

A tool which determines a disease activity biomarker based on NAS component scores (steatosis, hepatocellular ballooning, lobular inflammation) and fibrosis stage in biopsies in MASH clinical trials. The tool is an aid to a single central pathologist that is to be used for enrolment/inclusion of patients into clinical phase 2 and phase 3 trials in MASH as well as for the evaluation of the study outcomes (primary or secondary) in case this is intended to be based on histology evaluation. 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.

EMA/CHMP qualification opinions or qualification advice of novel methodologies for medicinal product development are provided without prejudice to any requirements related to other applicable legislation (e.g. MDR/IVDR and AI Act).

Executive Summary

The Applicant, Path AI, Boston, USA, proposes to qualify "AIM-NASH" (Artificial Intelligence-Based Measurement of Non-alcoholic Steatohepatitis Histology), to support the interpretation of MASH liver biopsies. The current method by which pathologists apply MASH histologic scoring systems is subjective and prone to variability; the machine learning (ML)-assisted approach aims to improve accuracy, repeatability and reproducibility.

The proposed context of use was initially proposed as follows:

"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 Non-alcoholic 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. 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."

After evaluation of the submitted material and discussion with the Applicant, the following context of use statement has been agreed:

" A tool which determines a disease activity biomarker based on NAS component scores (steatosis, hepatocellular ballooning, lobular inflammation) and fibrosis stage in biopsies in MASH clinical trials. The tool is an aid to a single central pathologist that is to be used for enrolment/inclusion of patients into clinical phase 2 and phase 3 trials in MASH as well as for the evaluation of the study outcomes (primary or secondary) in case this is intended to be based on histology evaluation. 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.

Consequently, the method can be used for all MASH clinical trials in which the histologic evaluation of liver tissue is used as part of the inclusion criteria, and/or efficacy evaluation.

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

The request for qualification has been submitted as a follow-up procedure of a previous Qualification Advice in 2021 which discussed the proposed development. The Applicant has finalised the programme and presented the results. The submission was received in June 2023.

The Applicant has presented the results of all studies within the development, including the description of the technical aspects comprising equipment, data platforms, proposed workflows, and other infrastructure elements.

The model development followed an iterative process with several phases of development. This 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 a so-called "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 again, predefined acceptance criteria assessed. The models met the pre-defined acceptance criteria on the "held-out test set". The model pipeline was then considered to be locked, and validation proceeded.

ML-based image segmentation models were developed for identifying image artefacts and key histologic features of MASH on H&E and trichrome WSIs based on annotations from different pathologists. 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 convolutional neural network (CNN) with stochastic minibatch gradient descent using the ADAM optimizer to produce pixel level predictions of NAS components (steatosis, lobular inflammation, and hepatocellular ballooning, fibrosis). Once the overlays were 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 MASH scoring models employing GNN were trained using a subset of the training dataset for the image segmentation models.

The presentation of the results allows the conclusion that the Applicant has followed the pathway discussed during the prior qualification advice which was considered acceptable.

The validation steps undertaken comprise the "stand-alone analytical verification" (with the "held-out test set"), the "integrated analytical verification," which tested the integration of the algorithm onto a digital pathology platform, the "validation of the AISight clinical trials platform", which in principle evaluated the similar performance of digitalized slides versus glass slides, "validation of the AISight translational platform", the "analytical validation", which tested the AIM-NASH algorithm without supervision of a pathologist, the "overlay validation" which tested the accuracy of a special feature of the AIM-NASH algorithm indicating the regions of interest, or "overlays," to the pathologist, and the final "clinical validation" which tested the accuracy of the tool under the intended final conditions.

For almost all of these validation steps the method used was to undertake the scoring of MASH features (fibrosis, steatosis, lobular inflammation and hepatocellular ballooning) with three methods of evaluation: A "ground truth" (GT) which was determined with the generally recommended and widely used method in clinical trials to have a panel of two central readers with a third acting as arbitrator in case of disagreements, an individual manual read (IMR) with a group of selected pathologists, and finally the AIM-NASH determined rating. Weighted kappas (WKs) were then determined as an expression of variability, comparing the manual reads, as well as the AIM-NASH scores with the ground truth, and then compare the results of the two based on pre-defined non-inferiority margins.

The main results of the evaluations contained in the clinical validation indicate that the algorithm-based evaluation of histology as compared to individual pathologist reads (IMR) proved to be non-inferior for the features of fibrosis grading, and steatosis stage, and proved to be superior regarding the stage of lobular inflammation and hepatocellular ballooning. Superiority was therefore concluded for those features, which are known to possess the highest variability.

The Applicant has also presented literature which included the re-evaluation of several finalised trials with the tool, presenting the change of the overall results, which is considered a valuable exercise, and contribute to the overall validity exercise and demonstrates that results become clearer, without introducing bias.

The tool has been "locked" regarding further training input. The Applicant has stated their intent that in case "major" changes need to be implemented (e.g. a relevant amount of new training data are included, relevant technical changes occur), further validation data will be generated. This is agreed and optimisation of the model is highly encouraged, but major changes might require re-qualification of the tool. The qualification is therefore valid only for the instrument in its current 'locked' state, without any "major" change to be implemented. Since from a regulatory perspective there for the time being is no post-qualification procedure in place such as e.g. a "Qualification Variation", the further development of the tool (if any), including minor changes, will have to be documented and presented in relation to any marketing authorisation application (MAA) for a medicinal product that has used the tool within their phase 2 and/or phase 3 development. The owner of the AIM-NASH tool as well as potential future Applicants for medicinal products will need to take account of this and provide appropriate documentation/justification as part of the MAA dossier.

However, since the general properties of AI-based tools as well as potentially multiple "major" changes could render assessment of additional validation data within a MAA impractical, the need for "re-qualification" should be considered if such changes are introduced and assessment within a MAA submission would need to be discussed with CHMP before submission.

Since the nomenclature for NASH and NAFLD has recently undergone major re-labelling to MASH and MASLD, respectively, the Context of Use statement has been updated accordingly. However, since the qualification documents were prepared and submitted prior to the change in terminology, apart from the Executive Summary, the old terminology is still used throughout the document and as part of the tool name, or other labels. It has been discussed whether the new definitions have an impact on the validity of the tool, and it has been accepted that the change in nomenclature is not regarded to be a relevant factor for the validity of the results of the algorithm or the validation exercise.

Abbreviations:

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 Non-alcoholic 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, US
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
MASH	Metabolic dysfunction associated steatohepatitis
MASLD	Metabolic dysfunction associated steatotic liver disease
ML	Machine learning
NAFLD	Non-alcoholic fatty liver disease
NAS	NAFLD Activity Score
NASH	Non-alcoholic 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

General note

Of note, the nomenclature of Non-alcoholic fatty liver disease (NAFLD) and Non-alcoholic Steatohepatitis (NASH) has been changed in a multi-stakeholder, world-wide consensus process. The results of the nomenclature (and partly the definition) changes have only recently been published online (Rinella *et al*, 2023*). NASH is now termed Metabolic dysfunction Associated Steatohepatitis (MASH).

As the qualification was filed before the new nomenclature was accepted, the term NASH is still used in this document.

* Rinella ME, et al; NAFLD Nomenclature consensus group. A multi-society Delphi consensus statement on new fatty liver disease nomenclature. Hepatology. 2023 Dec 1;78(6):1966-1986. doi: 10.1097/HEP.0000000000000520. Epub 2023 Jun 24. PMID: 37363821; PMCID: PMC10653297.

Qualification Opinion agreed by CHMP

Based on the evidence presented in the qualification opinion request and in a discussion meeting, CHMP considers that Artificial Intelligence-Based Measurement of Non-alcoholic Steatohepatitis histology (AIM-NASH) tool can be used in MASH clinical trials as an aid to trial pathologists in scoring the liver biopsies at enrolment and follow-up visits.

Agreed Context of Use (CoU):

A tool which determines a disease activity biomarker based on NAS component scores (steatosis, hepatocellular ballooning, lobular inflammation) and fibrosis stage in biopsies in MASH clinical trials. The tool is an aid to a single central pathologist that is to be used for enrolment/inclusion of patients into clinical phase 2 and phase 3 trials in MASH as well as for the evaluation of the study outcomes (primary or secondary) in case this is intended to be based on histology evaluation. 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.

Rationale for AIM-NASH development

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. Inter-reader agreement assessed by Weighted Kappa (WK) statistics is only poor-to-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. 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, as well as the borderline and challenging nature of some trial biopsies.

AIM-NASH can be used to support pathologic interpretation of NASH liver biopsies by accurately, consistently, 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 machine learning (ML) -assisted approach is expected to be more accurate and, most importantly, more reproducible. Therefore, AIM-NASH has the potential to be used to aid clinical trial enrolment and histologic endpoint assessment, providing value for accelerated and traditional NASH drug approval pathways.

Short description of the tool

AIM-NASH is an AI-based measurement (AIM) tool that provides Clinical Research Network (CRN) NAFLD Activity Score (NAS) component grades and fibrosis stages with pathologist review of the liver biopsies.

Technical aspects

Biomarker Source

The source of the biomarker is a formalin fixed paraffin embedded 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. Glass slides are scanned at a resolution of 40X.

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 septa, pathological periportal fibrosis, and fibrotic septa, but with a single output score (fibrosis stage). Fibrosis stages range from 0-4, where stage F0 indicates absence of fibrosis in the liver while stage F4 indicates cirrhosis.

Technical Platform

The technical platform used to measure the biomarker consists of the following elements:

1. Qualified WSI Scanner at Trial Site/ Trial Laboratory: 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. PathAI provides an Instructions for Use (IFU) that details use of the tool.

Information Security Management System (ISMS)

The technical platform development and deployment are compliant with the GDPR policies.

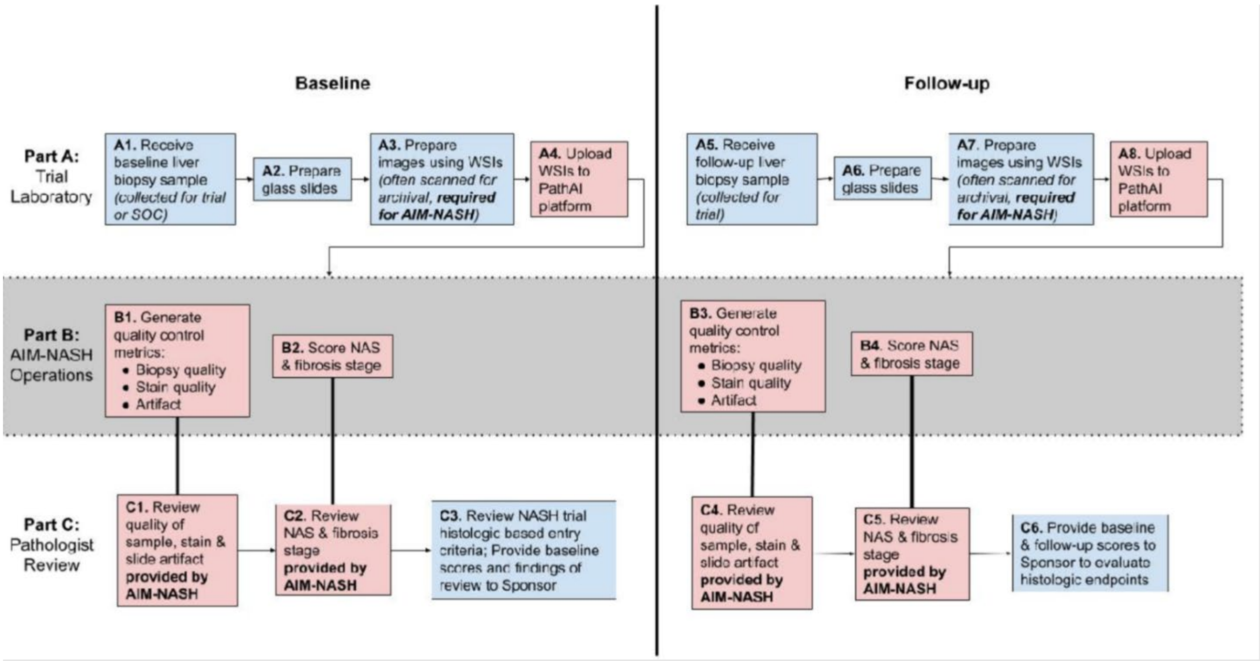
Setting and Workflow

The proposed context of use clearly states that the instrument is intended to be used in the clinical trial setting only. It was the aim of the development that the integration of the tool would not impact the overall trial workflow in significant manner. Sampling, sample preparation, staining and scanning will not be different from what is currently included in the conduct of clinical trials in the field, and the involved pathology labs are requested to be CAP/CLIA (College of American Pathologists/Clinical Laboratory Improvement Amendments) accredited. The qualifications of the pathologists involved in the validation studies has been well documented by the Applicant and has – besides documentation of the formal qualification – included a wide range of professional experience (e.g. years of experience ranging from 4 years to more than 30 years). This is agreed.

With respect to the use of the tool after validation, the Applicant states that the same evaluation exercise will be used to recruit histopathologists who are board certified and licensed in their respective countries/regions. Additionally, labs qualified in their respective countries/regions per applicable regulations and standards will be used. The statements are considered reassuring as such. Since, however, all validation data have been generated in the US, the Applicant is recommended to include pathologists from other regions in case any further evaluations/validation are taking place (i.e. when major changes occur that make additional studies or re-validation necessary).

The current routine workflow includes the generation of digital slides (termed WSI=whole slide images) which are used (mostly) for archival purposes (the upload of the WSIs to the PathAI platform, however, will be a new workflow element). The workflow items mainly influenced by the tool will be the evaluation of the quality of the sample, of the staining, and the artefact evaluation, as well as the final scoring of the NAS elements and fibrosis. For these items, a ”double” evaluation will be included which is done by the AIM-NASH operations in automated manner, as well as by the pathologist (local or central). Please see Figure 1 below.

Figure 1 : AIM-NASH in the NASH Clinical Trial Workflow



Baseline
Follow-up
Part A: Trial Laboratory
A1. Receive baseline liver biopsy sample (collected for trial or SOC)
A2. Prepare glass slides
A3. Prepare images using WSIs (often scanned for archival, required for AIM-NASH)
A4. Upload WSIs to PathAI platform
A5. Receive follow-up liver biopsy sample (collected for trial)
A6. Prepare glass slides
A7. Prepare images using WSIs (often scanned for archival, required for AIM-NASH)
A8. Upload WSIs to PathAI platform
Part B: AIM-NASH Operations
B1. Generate quality control metrics:
• Biopsy quality
• Stain quality
• Artifact
B2. Score NAS & fibrosis stage
B3. Generate quality control metrics:
• Biopsy quality
• Stain quality
• Artifact
B4. Score NAS & fibrosis stage
Part C: Pathologist Review
C1. Review quality of sample, stain & slide artifact provided by AIM-NASH
C2. Review NAS & fibrosis stage provided by AIM-NASH
C3. Review NASH trial histologic based entry criteria; Provide baseline scores and findings of review to Sponsor
C4. Review quality of sample, stain & slide artifact provided by AIM-NASH
C5. Review NAS & fibrosis stage provided by AIM-NASH
C6. Provide baseline & follow-up scores to Sponsor to evaluate histologic endpoints

The proposed workflow includes that the pathologist will assess sample quality according to the clinical trial protocol, and if deemed acceptable, the pathologist will review the H&E and trichrome based NAS component scores, and fibrosis grade generated by AIM-NASH. If the pathologist accepts these scores (within +/- 1 point per individual feature), they will record their agreement and sign-out the case. In case there is a 2-point deviation, these slides/WSIs will be rejected and send for consensus review. In rare cases where the primary pathologist (trial pathologist) and consensus pathologist (secondary pathologist) do not come to an agreement for scores, the primary pathologist will enter their scores which are then considered final.

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. A concern with this approach would be that in a trial slides may be scored primarily based on AIM-NASH and rejected slides scored by two human pathologists. The owner of the AIM-NASH tool has foreseen monitoring of the “rejection cases”, which have – up to now – been recorded in negligible numbers only.

It is advised for medicinal product developers using this tool, to plan sensitivity analyses to evaluate if drug effects may be different based on primarily AI (so only AIM-NASH scores are considered) versus human slide readouts for the rejected slides.

Additionally, algorithm overlays (see below) can be toggled off and on to facilitate review based on pathologist preference. The proposed workflow is considered acceptable, and it is understood that the procedures following rejection are subject to individual protocol definitions.

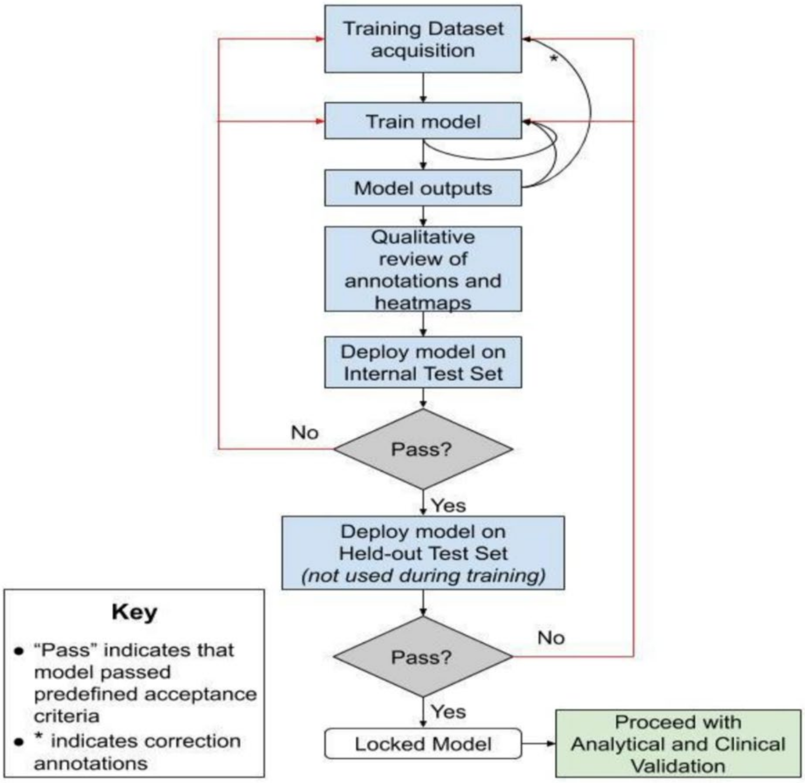
In other words, the AIM-NASH tool is proposed as a supplement to pathologist review and not as a substitute. The tool is always intended to be used in conjunction with the assessment of a qualified liver pathologist.

The workflow also includes two possibilities to upload the final evaluation, either by using the AISight Clinical Trials Platform with manual selection of the WSIs and subject metadata, or to use the "bulk ingestion mechanism" allowing the uploading of the WSIs and metadata directly to the S2 bucket on the AWS (used as "storage system").

Model development

The model development followed an iterative process as given in the following figure:

Figure 2: AIM-NASH Iterative Model Development


Training Dataset acquisition
Train model
Model outputs
Qualitative review of annotations and heatmaps
Deploy model on Internal Test Set
Pass?
No
Yes
Deploy model on Held-out Test Set (not used during training)
Pass?
No
Yes
Locked Model
Proceed with Analytical and Clinical Validation
*
Key
• "Pass" indicates that model passed predefined acceptance criteria
• * indicates correction annotations

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.

These two test sets in principle refer to a two-stage evaluation of different sets. However, since the test sets were also referring to different substances tested for the treatment of NASH (see table 2), the two sets do indeed cover a more diverse set of data, which is welcomed.

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 below).

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 table displays the overall dataset characteristics. The number of images used was 6227. Importantly, in addition to collecting annotations of the features relevant to the CRN scoring system, the model was trained to identify other features to increase the specificity of the model in recognizing the NASH CRN components.

Table 1: Dataset characteristics for H&E and trichrome segmentation model development

NAS Feature (H&E)	**Training Dataset**	**Fibrosis Stage (trichrome)**	**Training Dataset**
Number of Images, n	6227	Number of Images, n	6215
NAFLD activity score			
NAS<4, n (%)	1140 (18.3%)	0, n (%)	222 (3.6%)
N/A	1519 (24.4%)	1, n (%)	279 (4.5%)
Steatosis		2, n (%)	481 (7.7%)
0, n (%)	649 (10.4%)	3, n (%)	1453 (23.4%)
1, n (%)	3738 (60.0%)	4, n (%)	2173 (35.0%)
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%)

The training dataset seems to be sufficiently extensive, including slides from different trials testing different pharmacological entities. Inclusion of the non-NASH tissues is highly appreciated. However, certain disease features are poorly represented (e.g. F0/F1/F2 stage fibrosis, lobular inflammation grade 0 and steatosis grade 2 and 3) and it was questioned if the model is trained enough to recognize these features. The Applicant recognises the risk of the "unbalanced" dataset in model development and attempted to minimize it in two ways. During tissue model development, augmentations were applied to each sample, effectively increasing the number of samples even for rare classes. During development of the scoring model, samples were up weighted of rare classes. Specifically, samples were not picked with uniform probability, but instead inversely proportional to the logarithm of their relative frequency.

Moreover, the Applicant states that confirmation of satisfactory performance of the tool in currently underrepresented grades/stages is included in their monitoring plan. Data will be collected through agreements with biopharma partners and will be used to determine, implement, and monitor any necessary preventive and corrective actions. The collected data will be reviewed to address questions such as whether there is: an impact to the benefit/risk of the tool; a need to update design information or context of use; improvement necessary to the usability, performance, or safety of the tool. Monitoring reports will be created and reported to the European Medicines Agency (EMA) as part of MAA dossiers which present AIM-NASH derived evidence. The objectives of the monitoring plan will include detection of population changes; monitoring underrepresented score categories in current clinical trial populations; and tracking AIM-NASH 1-point and 2-point discordance rates. To achieve these objectives, cases will be collected from both the screening and follow-up timepoints, spanning all score categories and levels, and compared to consensus ground truth scores. Further, for the extremes (fibrosis 0 and 1 and lobular inflammation 0), additional cases will be collected, and any discrepancies will be evaluated. Discordance rates will be collected, and significant shifts will be investigated.

Data collected during monitoring will be used to continuously inform PathAI of the performance and continued safety and effectiveness of the tool. Changes to the safety and effectiveness of the tool are intended to be communicated to the EMA by the Applicant. Any necessary verification and/or validation of the tool will be documented. Since from a regulatory perspective for the time being there is no post-qualification procedure in place such as e.g. a "Qualification Variation", the further development of the tool (if any), including minor changes, will have to be documented and presented in relation to any marketing authorisation application (MAA) for a medicinal product that has used the tool within their phase 2 and/or phase 3 development. The owner of the AIM-NASH tool as well as potential future Applicants for medicinal products will need to take account of this.

These lifecycle-management efforts are acknowledged and agreed.

The number of, enrolment characteristics, phases and other features of the trials on which the development of the models were trained are given in the following table.

Table 2: Datasets used for developing the 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

The variety of different trials, and substances, and the inclusion of some other diseases (hepatitis B and primary sclerosing cholangitis) has been discussed and found to be satisfactory.

The Applicant collected annotations from a large panel of expert pathologists for which the following qualifications were requested: Board certification in pathology, as well as liver pathology subspeciality as evidenced by liver pathology fellowship training and/or significant ongoing clinical experience. Deep convolutional neural networks were trained, using the annotations generated by the 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 annotations (as defined by incorrect identification of substances by internal expert pathologists) 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. On trichrome age segmentation models were trained to detect fibrosis subtypes, including large septal, portal area fibrosis, subcapsular fibrosis and fibrotic septa. 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.

The large number of pathologists giving input is thought to have prevented the models from overfitting to a single pathologist's interpretation of the histology by the Applicant. The number of pathologists was 76, providing a total of 116,346 annotations. This is considered satisfactory.

Model Architecture, Training and Inference:

For the models 1-3, annotations were grouped into classes as appropriate and then used to generate training sets of image patches on the order of 500,000 samples. The features annotated not only comprised the NAS features of lobular inflammation, hepatocellular ballooning and macrovesicular steatosis, but also microvesicular steatosis, portal inflammation, interface hepatitis, and normal hepatocytes (for the H&E slides), and pathologically thickened fibrous septa, pathological partial and perisinusoidal fibrosis, normal portal areas (small), large normal septa (large septa) and subcapsular fibrosis (for the trichrome slides). These patches were used to train a deep CNN with stochastic minibatch gradient descent using the ADAM optimizer to produce pixel level predictions of NAS components (steatosis, lobular inflammation, and hepatocellular ballooning). Models 1-3 were 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.

Models 1-3 comprised Model 1 for artefact detection, and Models 2 (H&E Tissue Model), 3a (Trichrome

Tissue Model) and 3b (Trichrome Large Septa Model). 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.

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. The overlays are pixel-level model predictions. 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.

For the Models 4 and 5, ML CRN scoring models were developed for scoring the NAS components from H&E WSIs (Model 4) and fibrosis on trichrome WSIs (Model 5). 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 10 NASH pathologists from the PathAI expert contributor network. 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. 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 3 and Table 4.

Table 3: 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.%)

Feature	Training Dataset	Internal Testing Dataset (consensus)	Total
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 4: 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%)

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 modelled by a graph structure, such as fibrosis architecture. 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.

Models 4-5 comprise the components of the NAS scoring system with steatosis, ballooning, and lobular inflammation in the Models 4a-c, and fibrosis in the Model 5.

In general, the approach in model training that includes 5 models that cover recognition/annotations of histological features and eventually their scoring is considered adequate.

The Applicant concludes with a listing of the limitations of the model development, including the missing encounter of artefacts, colour markers, or unknown artefacts, the missing inclusion of non-liver type tissue, potential staining and scanner variations not encountered during development, missing of slides with very little usable tissue, and missing of biopsies being broken into multiple fragments. The relevance of these missing items may, however, regarded to be relatively low, since the generation of the AIM-NASH is not a fully automated process, but always undertaken with input from qualified pathologists.

The Applicant also states that any changes, including bug fixes or system patches, to locked and validated AIM-NASH will be 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 international 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 re-validation.

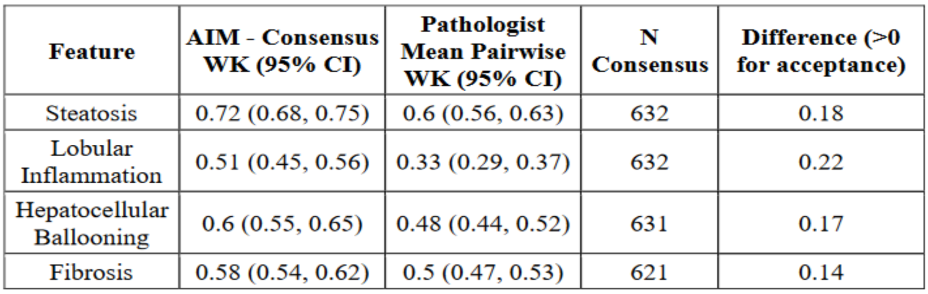
The Applicant has also provided a "revision level history" table for AIM-NASH in one of the appendices of the submission, which currently comprise 3 changes of the model (Versions 1.2.1, 1.3.1, and 1.4) which all have been classified as "low-risk change". Part of these conclusions are based on a so-called "regression testing" on a limited number of slides (Versions 1.2.1 and 1.3.1), or features are not used for the enrolment of patients or evaluation of efficacy. This is agreed.

Internal Test Set

In the process of model development, separate steps with different datasets, both not used in training, are utilized for verification purposes: Internal Test Set and Held-out test set (also called Standalone analytical Verification). Once satisfactory performance on training data was achieved, the models were deployed on the Internal Test Set and predefined acceptance criteria were assessed. Please see Table 3 for the details on the samples included in this set.

The Internal Test Set was a Phase 2 NASH trial evaluating a novel insulin sensitizer. Six-hundred and thirty-two (632) cases were utilized, and the dataset contained WSIs from a population with a range of fibrosis stage and NAS≥4 with a score of at least 1 in each component of NAS. Slide level scores for NAS components and fibrosis were collected from 3 expert liver pathologists. Agreement of AIM-NASH read-outs with mean consensus pathologist reads was assessed using linearly WK statistics. For reference, pairwise pathologist agreements were also computed. Even though the acceptance criteria were not documented prior to the internal testing, the results met the standard acceptance criteria of 0.1 non-inferiority, according to the Applicant. Indeed, for all histologic features, agreement of AIM-NASH read-outs with consensus reads was greater than the pairwise agreement between pathologists performing manual reads for internal test set verification.

Table 5 : Agreement of AIM-NASH and Pathologist Mean Pairwise Comparison from Internal Test Set Verification



Feature	AIM - Consensus WK (95% CI)	Pathologist Mean Pairwise WK (95% CI)	N Consensus	Difference (>0 for acceptance)
Steatosis	0.72 (0.68, 0.75)	0.6 (0.56, 0.63)	632	0.18
Lobular Inflammation	0.51 (0.45, 0.56)	0.33 (0.29, 0.37)	632	0.22
Hepatocellular Ballooning	0.6 (0.55, 0.65)	0.48 (0.44, 0.52)	631	0.17
Fibrosis	0.58 (0.54, 0.62)	0.5 (0.47, 0.53)	621	0.14

Standalone analytical Verification (SAV)

The standalone analytical verification is in principle still part of the Model development (see above Figure 2) and has thus to be considered a prerequisite for model lock, and continuation with the planned analytical and clinical validation programme. The so-called "Held-out" test set was used for this part of the verification (using the AWS platform), and samples were taken from two phase 2 studies: ATLAS study, Gilead, and a semaglutide study, Novo Nordisk. These held-out sets were not used for model training. The Gilead study included 900 slides (each for H&E and trichrome) from patients with F3-F4 fibrosis, and the Novo Nordisk study included 530/532 H&E or trichrome slides from patients with F1-F3 fibrosis. Three pathologists generated the manual scores. In total, 250 cases were selected in order to include about 50 cases for each of the fibrosis stages (F0-F4).

The final slide distribution (according to the manual reads) reflected all grades/stages of all features of NASH appropriately, except for lobular inflammation where only 10 cases with grade 0, and only 2 cases with grade 3 were included.

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 (for the justification of these criteria, see also below in the Analytical Validation section).

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 WKs generated during SAV from pathologists.

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. The following results were achieved:

During the evaluation, changes to the algorithm were introduced going from Version 1.0.0 to Version 1.1.0. This change was assessed as being of low risk, and no need was stated to repeat the exercise. However, the new version was to be used in the following step, the "Integrated analytical Verification (IAV)".

The Applicant concludes that the algorithm results generated during SAV on the AWS development environment met the prespecified acceptance criteria.

Table 6: 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)

Integrated analytical Verification (IAV)

The objective of this part of the validation (verification) is described as follows:

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 this exercise, the acceptance criteria were set as "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."

For this exercise, the locked model was deployed on the held-out test subset, and it was required that the results from this activity must match the results of the SAV. 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.

Results:

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.

The Applicant concludes 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.

Validation of the AISight clinical trials platform

The AISight Clinical Trials platform (v3.3.1) is a RUO (research use only) 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 to be used by 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, including quantitative assessments of histological features, cell density and spatial relationships.

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

The evaluation of agreement between WSI and manual reading on glass slides was part of the proposals made during the 2021 advice and was supported.

The Applicant 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.

Ground truth (GT) reads were done by 3 board-certified hepatopathologists using glass slides with each of the slides read by 3 pathologists. The GT score was computed as the median of all 3 scores. It is not clear why GT was determined not as a consensus read, but rather the median of all 3 reads, but for the purpose of the exercise this can be agreed.

The "test sets" were read by 3 board certified hepatopathologists first on the WSIs, and, after a "wash-out" of 2 weeks on the glass slides.

The slide set was to consist of 160 cases, and hence of 320 slides (H&E and trichrome). Two thirds of the cases were to be 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 was from NAFLD patients (without NASH) and other (non-NAFLD) liver indications, including but not limited to hepatitis B, hepatitis C, active hepatitis, viral hepatitis, cirrhosis and cholestatic liver disease. 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.

The pathologists were blinded to each other's assessments and to their own assessments from the different modalities.

The primary endpoint was 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 for NASH clinical trial enrolment, 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 appropriateness of this endpoint was questioned. The Applicant argues in favour of this endpoint, as it takes into account scoring of different NASH components and any additional findings that the pathologists might have identified during a clinical trial. This is understood, however the comparison of GT to Glass/WSI study pathologists on the level of individual NASH components scores was also considered of interest. In response, the Applicant provided the requested data (please see results).

The null hypothesis was 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. The secondary endpoint was the evaluation of the 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 based on 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 was to be computed by averaging the WK for the 3 pathologists. Bootstrap 95% confidence intervals were to be provided on the overall as well as per pathologist linearly WK. These concordance estimates were to be compared to the published range in the following table:

Table 7: 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.

The sample size was on one hand based on the College of American Pathologists (CAP) guidelines which recommends a minimum of 60 cases when evaluating the diagnostic performance of digitized slides compared to glass slides. 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 therefore 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.

Results:

Finally, 318 slides (of 159 cases) were enrolled into the study, with 186 slides from the Madrigal (resmetirom) phase 2 study, and 36 from the phase 2 NorthSea Icona Trial (carboxylic acid), both applying the "usual" inclusion criteria regarding NAS score, but with fibrosis ranging from F1 to F3. The validation further included 18 cases from Precision of Medicine. These cases were all non-NASH samples that included normal liver and other liver disease indications (e.g. viral hepatitis) that could be encountered in a NASH clinical trial. The other slides used for the study were from two different NASH trials, including both screen failures and enrolled biopsies. Out of the 159 (NASH) cases, the distribution of different NASH features was not equal, with the majority of cases having higher grades of the NAS features and having fibrosis stages 1-3 (reflecting the inclusion criteria of the trials of the NASH cases). The various sources of the slides are considered to sufficiently cover the potential variety of underlying pathology. Although a separate evaluation of the influence of e.g. the quality of stains, section thickness, age and other parameters of the glass slides was not performed, the comparison made was considered valid, since the variability in these parameters was considered similar for the glass as well as the WSIs. However, the Applicant is recommended to address different quality attributes of the glass slides in the future, when additional scanners, other than the Aperio AT2 are validated.

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). 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).

The evaluation of agreement for the single pathologists revealed that 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).

The following table shows the secondary analysis of the WKs, which obviously compare well with the published data as given above. If the more "optimal" values of the Kleiner (Kleiner 2005) study are considered (WKs were 0.83, 0.60, 0.66, and 0.85 for steatosis, lobular inflammation, ballooning, and fibrosis, respectively), the WKs appear still numerically superior, however, without relevant differences for steatosis and fibrosis, but still high for lobular inflammation and ballooning.

Table 8*:* 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)

Additional evaluations were performed by the individual pathologists (not shown) which reflect the variability between the 3 pathologists, demonstrating (similar to the individual primary endpoint evaluation) that pathologist A performed “best” followed by pathologist C, followed by pathologist B.

Also, agreements for all component scores and overall NAS, by WK as compared for glass and digital slides vs. GT were presented as requested. The agreements are equivalent for glass and digital vs. GT, except for hepatocellular ballooning, where the digital agreement with GT was higher than for glass, with non-overlapping confidence intervals.

The Applicant concludes that the study supports the conclusion that the AISight Clinical Trials platform is non-inferior to the glass read, and that the adequacy of the WSIs scanned by the Aperio scanner can be adequately used instead of glass slides. This conclusion is supported. It is to be noted that the method implemented in the digitalization of slides is currently restricted to the methods used in this trial and, as indicated in the CoU statement, the use of the Aperio AT2 scanner only. The Applicant is recommended to assess the influence of the properties of the Aperio AT2 regarding the techniques for digitization of slides (data formatting, image data compression, and storage of meta-data etc) when it comes to validation of additional scanners.

Validation of the AISight Translational Platform

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.

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

The design of the study, the dataset and endpoints were similar to those described above for the AISight Clinical Trials platform validation. Therefore, only the results for the AISight Translational platform will be presented here.

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). 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).

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. 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).

The following table shows the secondary analysis of the WKs. The WK values are generally similar to what has been shown for AISight clinical trials platform, with the exception of lobular inflammation and hepatocellular ballooning for which WKs were substantially lower but still in the range of reported literature. The differences in absolute WK values between platforms are within the expected range of variability demonstrated by different NASH experts as reported in the literature.

Table 9: 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)

Analytical Validation

General aspects

The purpose of this study was 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. Simply speaking, this part of the validation exercise tests the accuracy, reproducibility and repeatability of the AIM-NASH score generation without supervision and input of pathologists (the full workflow including pathologist review and input is tested in the "clinical validation" (see below).

Analytical validation of AIM-NASH was performed based on glass slides provided by sponsors and selected from completed phase 2 (Bristol Myers Squibb FALCON 1 trial NCT03486899 and FALCON 2 trial NCT03486912 with Pegbelfermin) and a phase 3 NASH trial (Intercept Pharmaceuticals REGENERATE trial, using obeticholic acid). The included population was a stage 1-3 fibrosis population in the REGENERATE, a fibrosis stage 3 population in the FALCON 1, and a fibrosis stage 4 population in the FALCON 2 trial. These cases/slides were independent from those used in the training and previous validation exercises. The number of cases intended to be evaluated was n=600 for accuracy, and n=150 for repeatability and reproducibility.

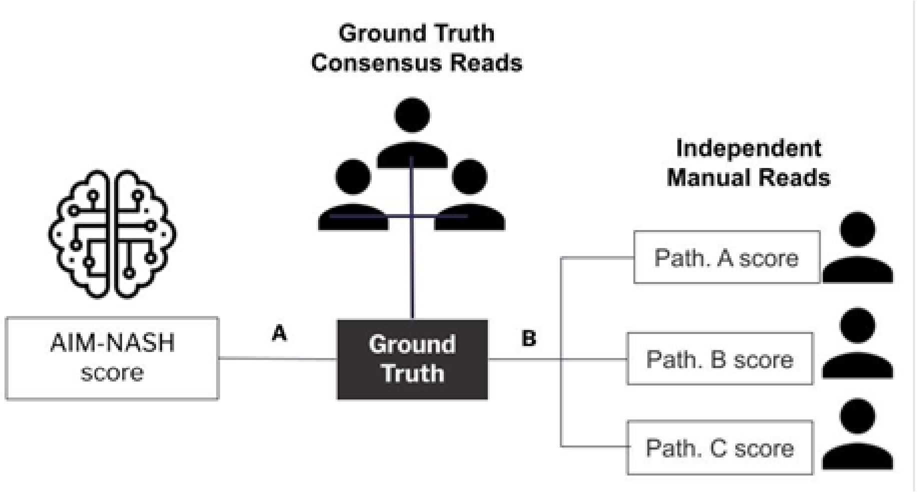
All slides for accuracy study were scanned at the Covance Indianapolis site (now LabCorp) on Aperio AT2 scanner at 40x magnification following established lab Standard Operating Procedures (SOPs). After scanning was complete, Covance lab technicians uploaded the WSIs onto AISight Clinical Trials platform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

There were three primary objectives of the trial:

- **Accuracy**: To evaluate for non-inferior agreement, linearly WK for AIM-NASH scores vs. GT was calculated and compared that to mean pairwise WK for Independent Manual Reads (IMR; with at least 3 pathologists) vs. GT for all 4 NASH components. Please see figure below for the schematic view of accuracy testing.

The GT was established by two panels of 2 expert liver pathologists with a third pathologist serving as tiebreaker, and the tiebreaker blinded against the disagreement. The pathologists were chosen based on their previous MASH experience and results of previously completed work in the MASH field for PathAI. All slides from AV (analytical validation) and CV (clinical validation) were split between the two panels, so that each panel read about half of the overall dataset (AV and CV slides combined; see also below for the CV).

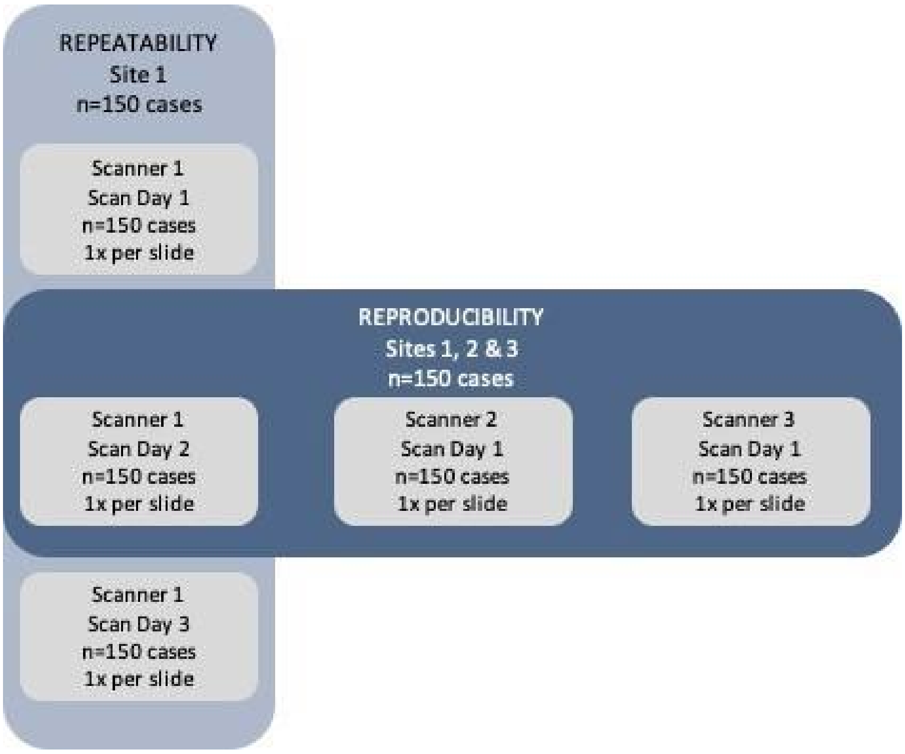
Figure 3 : Accuracy Study Design



- **Reproducibility** (inter-operator and inter-scanner variability): 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 the Applicant aimed to show statistical superiority above this threshold. A subset of 150 slides were used for this part.

- **Repeatability** (intra-scanner, time-dependent variability): 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 the Applicant aimed to show statistical superiority above this threshold. For this evaluation, also a subset of 150 slides were used. Please see figure below for the schematic view of repeatability and reproducibility testing.

Figure 4: Repeatability and reproducibility study design



As secondary objectives, accuracy was evaluated with clinical subsets (different visit such as baseline, post-baseline etc) for F4 or other, F0-1 or other, NAS≥1 or other.

Exploratory objectives were defined to provide summary statistics for accuracy reproducibility and repeatability reporting distributions of trial origin, baseline co-morbidities, NASH treatment, and other subgroups.

Methodology

For the sample size determination, the following methods were used:

Given a range of scores for each NASH component, inter-pathologist WK based on literature, and non-inferiority margin of 0.1, both the upper bound of inter-pathologist WK (Target) at 90% power and lower bound (LB) of WK between AIM-NASH model and consensus evaluated during internal testing studies at alpha of 0.025 were estimated based on the parametric model. Simulations were run to find the smallest N for which LB > Target - 0.1 passed. Based on the CRN literature (Kleiner et al. 2005, Davison 2020), different inter-pathologist WKs were expected slightly below the ranges in the Kleiner et al study. The sample size of 600 was selected since this was assumed to provide the most conservative estimate for the component with the most variable inter-pathologist WK - hepatocytic ballooning with a WK of 0.5.

The non-inferiority margin of 0.1 for Linearly WK was chosen based on the following argumentation:

1. Qualitative interpretation of WK: Linearly WK is a measure of ordinal class concordance where a value of 0 represents chance and 1 represents perfect agreement. WK 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 the chosen non-inferiority margin of 0.1, WK represents half-width of these qualitative bins.

2. Range of observed inter-pathologist WKs. Pairwise inter-pathologist WKs were measured for the held-out test set, across three qualified, expert pathologists (N=220 trichrome, N=231 H&E). WK 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 WK conforms to these ranges of inter-pathologist concordance for fibrosis, ballooning, and steatosis.

The given justification for the non-inferiority margin of 0.1 had already been discussed within an initial advice procedure during which the similar justification was given. The justification based on the categorisation of WKs, and the ranges observed were considered acceptable

However, improvement compared to published values intra- and inter-rater variability was tested across all score components during reproducibility studies, where the endpoint requires greater than 85% agreement across scans and scanning operators/sites to demonstrate superiority compared to relevant published WK values.

Intra- and inter-reader variability reference ranges were based on the published literature, and specifically the study by Davison et al. (Davison 2020). While it may go undisputed that the reported variability using inter- and intra-pathologist WKs may be a "best standard" for the CRN-network based evaluations (Kleiner et al 2005, and 2019), the frequently cited study by Davison et al appears to be rather a "worst case" scenario, and appears to be "a priory" inferior to the PathAI contributor network WKs that were used for the sample size estimation and the underlying simulations.

For the discussion about the "clinical relevance" of the results achieved, the Applicant was requested to not only make formal comparisons with the data reported from the Davison et al study but evaluate whether this study indeed presents a "worst-case" scenario, also because other literature references (also cited by the Applicant) have shown that a "better performance" of the "conventional methods" is possible (Sanyal A et al. 2021). The Applicant provided argumentation regarding the differences of the studies, referring to sample sizes, and – more importantly – the comparisons made. While the Davison paper compares pairwise WKs (of three pathologists), the Kleiner studies have reported average WKs across a whole group of up to nine pathologists. Contrary to this, the Sanyal study reported WKs across two different panels of three pathologists (with WKs between individuals). Also, the Applicant refers to the potential for higher variability considering the variety of data included in the AV and CV studies. In general, the comparisons in the AV and CV studies can stand on their own, and the "external comparisons" are only suitable to give a coarse orientation of the level of variance that is achievable with the AIM-NASH tool (see also Clinical Validation).

Results:

The final "study population" comprised of 165 subjects from the FALCON 1, 105 subjects from the FALCON 2, and 238 subjects from the REGENERATE trials, see below on dataset. Due to inclusion of baseline as well as (partly) post-baseline samples, this comprised finally 251 samples from REGENERATE, 217 samples from FALCON 1 and 139 samples from FALCON 2, and hence a total of 607 samples (the Briefing document erroneously reports 606, but the study report includes 607).

Dataset

Liver biopsy slides distinct from those used for training and validating the AIM-NASH model algorithm are obtained to support analytical validation.

Analytical validation of AIM-NASH was performed against glass slides provided by sponsors and selected from completed phase 2 trials (Bristol Myers Squibb FALCON 1 trial NCT03486899 and FALCON 2 trial NCT03486912) and a phase 3 NASH trial (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. 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.

Out of the 607 enrolled slides, less than 4% of the slides had missing final GT score due to various reasons. 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.

The distribution of the slides (according to the GT rating) was deviating from the planned "equal" distribution according to the staging and grading of the different aspects of evaluation. While for steatosis evaluation, at least of 15% share was achieved (for score 0 and 3), and for ballooning this was at least 13% (n=76), there was rather a low share of the lobular inflammation category 0 (2.5%, n=15), and for the fibrosis stage 0 (3.6%). For the total NAS score, this resulted in low shares of NAS of 0, 1, and 2, as well as for 8 (only 0.67%; n=4), but the dataset was well-represented around the NAS >=4 trial inclusion criteria for NAS 3, 4, 5 (18.72%, 22.9%, and 22.4% respectively).
A more equal distribution was achieved in the subset of the 150 (finally in fact 139-144 slide-sets were used) slides for the repeatability and reproducibility evaluation.

Table 10: 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 11: 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 12: 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 evaluation:

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. 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. The results are shown in the following table:

Table 13: 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)			

The primary analysis in consequence demonstrates success of the trial since all the lower bounds of the 95% CIs were >-0.1 in all cases. In fact, there was a slight numerical superiority for fibrosis and lobular inflammation, which might not be relevant. However, the AIM-NASH performed slightly inferior to the IMR evaluation for steatosis. Hepatocellular ballooning, which is the only item that demonstrated a significant superiority is the aspect of the CRN-scales that has always been characterised to be the most difficult. The WKs for AIM-NASH vs GT are considered of the most relevance in this analysis, as the tool is expected to replace GT in the clinical trials (i.e. panel of three pathologists will be replaced by the AIM-NASH tool used in conjunction with the review by one trial pathologist), not the individual pathologist scoring, therefore, even though primary endpoint comparisons are of interest and serve as proof of concept (reduction of inter-individual variability), they are of lesser relevance. A comparison to the literature-based (CRN-network) WKs shows that lobular inflammation performed in the range of what has previously been documented, while steatosis and fibrosis performed slightly inferior. Ballooning – at least for the AIM-NASH evaluation – was slightly better than the CRN-reported values. Compared with the "worst-case" WKs of the Davison study, however, all values appear to have improved.

In addition to this primary analysis, several aspects/factors of the overall evaluation were also similarly compared, of which the results are given in the following table:

Table 14*:* WKs for NASH aggregate scores

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)
	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)
	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)
	IMR vs GT	593	0.574 (0.513, 0.641)	
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)	

For none of these categories was the chosen non-inferiority margin of -0.1 missed, while the NAS ≥4 vs. <4, the NAS ≥4 and ≥1 for each component, and the NASH resolution categories indicated superiority.

The WKs provided for accuracy give an indication of agreement but do not allow an assessment of the type of disagreement. False positive and negative rates of AIM-NASH compared to the GT are needed to assess whether the disagreement is 'balanced' or whether there is a systematic deviation (this also applies to the clinical validation). The Applicant provided overall percent agreement, positive percent agreement and negative present agreement for AIM-NASH vs. GT and IMR vs GT for various measurements for analytical and clinical validation sets. Since these analysis for analytical and clinical validation produced similar results, they will be discussed only for clinical validation, as the most relevant. Please see Clinical Validation section.

Further secondary accuracy evaluations were also performed for each NASH component by time point (baseline or post-baseline). The time-points were generally in accordance with each other and showed accordance with the overall evaluation, including the superiority for ballooning. Similar general accordance could also be achieved for the evaluation of exploratory endpoints according to study except for the FALCON I trial, which performed inferior in the steatosis part (difference was -0.104 (-0178, -0.022). Similarly, the descriptive evaluations were also conducted for each of the NAS-scoring component scores (and fibrosis), which were again generally in line with the overall results, but non-inferiority not demonstrated consistently. The AIM-NASH performed, however, inferior regarding steatosis grade 1, but superior for all ballooning grades.

The Applicant correctly points out to the fact that the study was obviously not powered for the exploratory endpoints.

An exploratory analysis was also conducted regarding the exclusion of non-liver tissue containing slides which showed that the overall results were not changed.

Repeatability evaluation:

Repeatability endpoint for this study included scans from the same glass slides repeated over 3 non-consecutive days. The AIM-NASH tool was then deployed on each WSI and the agreement per histologic component was evaluated across WSIs. The results are shown in the following table.

Table 15*:* 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

* comparing to a performance goal of 85%

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.

Secondary repeatability endpoints:

Mean agreement of AIM-NASH scoring between inter-day timepoints were assessed across baseline and post-baseline time-points (including placebo and treatment groups) of the dataset. The mean agreement of all time points was significantly greater than 0.85.

Exploratory analysis also assessed each NASH component score level 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. However, this goal was not achieved for fibrosis score 1 (mean agreement rate of 0.808 (95% CI of 0.679, 0.912). The Applicant, however, claims that this agreement rate is still higher than reported intra-reader agreement for fibrosis overall in the literature (again referring to the Davison study) and, as pointed out previously, the study was not powered for individual score levels.

As a post-hoc evaluation, the agreement rates were also assessed for each of the components in the single studies contributing to the dataset. The 85% margin was achieved for all evaluations except for the FALCON 1 study steatosis, and the REGENERATE study Fibrosis evaluation (lower 95% CI values of 0.83 and 0.97).

A further post-hoc evaluation explored the AIM-NASH repeatability when deployed 5 separate times. This achieved perfect agreement of 1 with the lower bound of the 95% CI at 0.975 in all components.

Reproducibility evaluation:

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 agreement rates are shown in the following table:

Table 16*:* 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

As seen from the table, superiority could only be concluded for the ballooning component, while all other components achieved high reproducibility, with the lower bound of the 95% CI just above 80%. The Applicant refers to the values achieved for as documented in the literature (again referring to the study by Davison et al 2020, which showed reproducibility between 0.5 (for fibrosis) and about 0.6 (for the other components)).

The Applicant argues that the results achieved with AIM-NASH are still superior to what is known from various literature sources and that pre-analytical variability might have been a challenging factor here which could, theoretically, affect a tool’s intra- and inter-site measurements. These arguments can in principle be followed. Given the observations and results generated from this study and others, it is ideal to standardize lab processes as much as possible, including staining, scanning, and post scanning image QC processes, especially if multiple labs will be utilized in a trial or across phases, to minimize variability. Additionally, the Applicant argues that the pathologist will be reviewing the quality of the stain and scan and algorithm scores and has the ability to request a restain or rescan in a clinical trial workflow, unlike during analytical validation, where there was no pathologist review.

Furthermore, secondary and exploratory evaluations mainly did not meet the 85% threshold. However, the results were mostly in line with the primary evaluation regarding exploration/comparison of baseline and post-baseline categories. For the NASH score components (grades and stages), the results (based on descriptive statistics) were mostly in line for steatosis, inflammation, and ballooning, but showed inferior performance for the lower fibrosis grades 0 and 1, while the higher fibrosis scores were in line with the primary reproducibility evaluation. The Applicant points to the fact that these were based on particularly low numbers of samples (9 and 12 in the fibrosis grades 0 and 1, respectively, with a point estimate around 0.7). Post-hoc explorations included reproducibility regarding trial of origin, which was greatly in line with the primary evaluation, with some fluctuation between the trials and the exception of the REGENERATE trial fibrosis evaluation. This was attributed to the fact that the trichrome slides from that study represented a wide variety of stain quality, many being faded and several years old.

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 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) and can be considered reassuring.

Finally, the pairwise inter-reader agreements were calculated between IMR pathologists across all cases 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 pathologist readers who read at least 10 common cases). The results of this evaluation are shown in the following table:

Table 17*:* 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

In conclusion, 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.

The potential to minimize the inter and intra-reader variability was tested through the repeatability and reproducibility arms of this part of the validation exercise. For same site scanner repeatability, percent agreement was significantly higher than 85% for each of the four NASH components. Overall, reproducibility was near 85% for all components, but lower bounds of the confidence intervals fell below 85% for steatosis, inflammation, and fibrosis.

The Applicant finally concludes that 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 enrolment, monitoring, and therefore more accurate determination of histologic change over time for trial endpoints. These conclusions can mainly be followed.

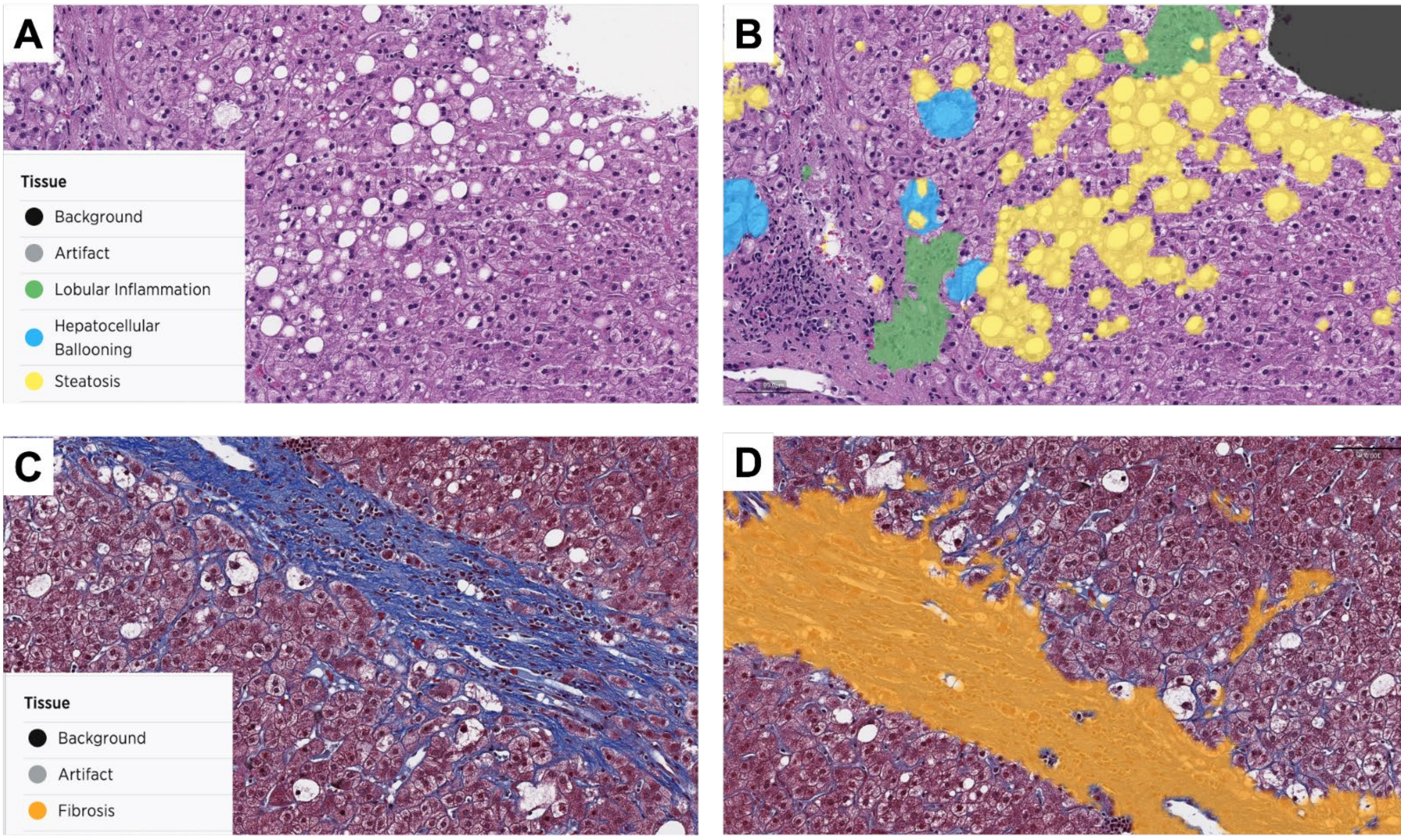
Overlay Validation

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 overlays give the pathologist the option of displaying the coloured overlays showing the CNN model's predictions of areas containing the histologic features of interest (i.e., steatosis, lobular inflammation, and hepatocellular 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.

In order to provide a visual impression of what an "overlay" is, the following figure is presented which gives representative "overlays" for the H&E and trichrome stained slides:

***Figure 5*:** Representative H&E and trichrome Overlays



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.

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 was intended to assure 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 Enrolment Task.

For the Frame Enrolment Task, a qualified PathAI Contributor Network pathologist who has demonstrated extensive experience reviewing NASH biopsies and was not involved in AIM-NASH model development was to perform frame selection. For each frame, the pathologist was 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

The enrolment evaluations were made without the use of the AIM-NASH overlays. The enrolment was aimed to fulfil specific feature "buckets" by an unblinded Path AI clinical data manager aiming at fulfilling the following requirements:

Table 18: 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	None	None	None	None	None

(8–16)	(8–16)	(8–16)	(8–16)	(8-32)	(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 enrolment 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 (performed by two panels of two pathologists, with a third tiebreaker pathologist when necessary; according to the rules used also in the other studies).

Once the frames were enrolled, they were sent out to 3 board-certified expert hepatopathologists, who were trained on the study protocol. The pathologists were blinded to each other's assessments and the enrolment 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.

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.

The primary objective/endpoint of this study was defined as to assess the accuracy based on acceptable level of sensitivity (true positive) and specificity (true negative) 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 liver biopsies stained with H&E or trichrome.

The exploratory analysis was to examine sources of variability between pathologists for features with more discordance between pathologists in identifying slides with that feature.

For the primary analysis, the overlay performance was considered acceptable if true positive (TP) success rate and false positive (FP) success rate were greater than or equal to 85%.

The rate of true positives (Question 1): was evaluated for:

- Macrovesical steatosis + fibrosis with the criterion
  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
- Artefact:
  The artefact is present in the frame and overlay is identifying artefact

The rate of false positives (Question 2): was evaluated for

– False Positive (FP; measure of specificity) success

Less than or equal to 20% of the total frame area is being overestimated by overlay

For each success criterion, 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 artefact 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 2.5th and 97.5th 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.

The sample size was determined based on the assumption of 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.

Results:

WSIs were selected from those also available for use in AV and CV for the AIM-NASH drug development tool (DDT) overlay validation. The slides were selected from four completed NASH phase 2b or 3 clinical trials (Table 19). Slides were selected such that they reflect a representative distribution of disease severity.

Table 19*:* 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	2	Pegbelfermin	Biopsy must be consistent with NASH

Squibb			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

The distribution of the frames regarding the different NASH features was not uniform but included rather low numbers of low scoring (especially in the slide level score, e.g. ballooning was 15.1% for grade 0, inflammation only contained 2.3% for grade 0, and steatosis only 5.8% for grade 0). Fibrosis stage 0 was present with 1.27%. The other scoring categories were well represented. A more "equal" distribution was, however, achieved for the frame distribution, but still the share of the lowest categories included 6.9% for lobular inflammation none, 6.25% for steatosis none, 10% for ballooning none, and 6.25% for fibrosis none. Slides and frames were quite equally distributed by the different sponsors.

Finally, a total of 160 frames per feature were evaluated.

As indicated above, 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 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).

The overall true positive success rates per overlay feature are shown in the following table.

Table 20: 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 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 also performed "worse" than the pathologists A and B for other features (also for the FP evaluation).

The acceptance criteria for FP (overestimation) success rate were met for all 6 feature overlays which is shown in the following table:

Table 21: 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)

For exploratory analysis variability between pathologists' identification of each feature was determined. 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 was determined at 89.2% for H&E artefacts, 80.0% for lobular inflammation, 99.4% for steatosis, 72.0% for trichrome artefacts, and 96.8% for fibrosis. The agreement for presence of hepatocellular ballooning was the lowest at 55.1% out of all features and therefore, sources of variability between pathologists for hepatocellular ballooning were further examined:

For the 65 frames where all 3 evaluating pathologists indicated presence of hepatocellular ballooning, the TP success rate was calculated. 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.

In their conclusions on the results of the study, the Applicant refers to limitations based on the missing of fibrosis stage 0 in the frames, the general fact that only frames, but not slides were examined, and the fact that a % area was to be examined also for the features of lobular inflammation and ballooning, for which usually not % area but number of foci are evaluated. However, the Applicant still overall concludes that the overlays are accurate as a spotlight to highlight the features of NASH slides.

It is obvious, and admitted by the Applicant, that only part of the primary objectives of the trial were met. The miss of the TP rate for the ballooning part of the evaluation, however, does not really come as a surprise, since ballooning has always been considered to be the most difficult and most variable part of the NAS-score assessments. However, formally, the study cannot be concluded to be successful. In this regard, the Applicant referred also to the known variability as documented in the literature (e.g. Brunt et al. 2022 reported a WK of 0.197 for ballooning). Additionally, the Applicant has evaluated the reasons for the different scoring of the pathologists for which it turned out that this pathologist identified a unique sub-set of cells as ballooned which the two other pathologists and the model did not, which may have resulted in the difference in assessment.

Considering the overall high rates as regards the rates of TP and TN, and the use of the tool as additional feature within the evaluation when a manual read is performed in addition to the AI-based evaluation, the deviation reported can be expected to have minor influence on the overall validity of the tool, and the argumentation of the Applicant was accepted.

Clinical Validation

Like the analytical validation study, the primary purpose of this study was 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. This part of the validation exercise tested the accuracy of the AIM-NASH score generation under inclusion of the full proposed workflow, i.e. with supervision and input of pathologists and was regarded the final step of evaluation.

The slides were previously scanned for the trial using Aperio AT2 whole slide scanners at 40x magnification.

Each case enrolled into this study was given a GT score which was evaluated by two panels of 2 expert liver pathologists with a third one acting as tiebreaker in case this was needed.

IMRs were performed by 8 qualified PathAI Contributor Network liver pathologists and each case enrolled received a minimum of 3 reads from the independent manual reads (IMRs).

The AI-assisted reads were performed by a set of 6 expert liver pathologists and one final AI-assisted read was assigned for each case. The pathologist workflow for reviewing AIM-NASH results involved sample quality assessment, evaluation for any potential, additional findings, and 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 was allowed to 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 (in case of disagreement by 2 or more of an individual score). Workflows were clearly defined and described in the study protocol.

The study plan defined for potential inclusion slides from the REGENERATE trial (phase 3 study for Obeticholic acid), the FALCON 2 trial (phase 2 study of pegbelfermin in the cirrhotic population), and a phase 2 study for semaglutide in the non-cirrhotic population. Both slides from patients with screen failures, as well as those from screening and (intermediate) endpoint evaluation, were included. The overall enrolment goal was 1424 cases, and final enrolment was 1501 cases.

During this study, the "OpenClinica" electronic data platform was used. This platform was used to enter data for glass slide reads in the platform because the AISight Clinical Trials Platform did not have any data entry capabilities at the time of these studies, and additional information data capture was necessary per study design (e.g., overlay utility questions for study pathologists being assisted by AIM-NASH).

Reassurance was received that the OpenClinica eDC platform was only utilized in the validation studies and will not be part of the final data management for the algorithm workflow for end pathologist users during NASH trial evaluations. AISight Clinical Trials Platform (the final platform) has the same data entry capabilities as the OpenClinica eDC platform and will be the only platform used for AIM-NASH in the intended use, i.e. in clinical trials.

The primary objective of the trial was to evaluate the performance of the AIM-NASH assisted pathologists. The primary endpoint was defined as to evaluate performance of the AIM-NASH tool in a clinical trial setting and demonstrate noninferiority for all assessment scores in terms of accuracy against a panel of qualified histopathologists. Similar to the analytical validation, the aim was to show non-inferiority with a margin of 0.1 for the inter-pathologist WKs. Please see justification for these criteria in Analytical Validation section.

Given a range of scores for each NASH component, inter-pathologist WK based on literature, and non-inferiority margin of 0.1, both the upper bound of inter-pathologist WK (Target) at 90% power and lower bound (LB) of WK between AIM-NASH model and consensus evaluated during internal testing studies at alpha of 0.025 were to be estimated based on the parametric model. Simulations were run to find the smallest N for which the below test passed:

LB > Target - 0.1

As mentioned, the protocol determined the necessary sample size (enrolment goal) as 1424 cases.

The secondary endpoints were defined as to assess the 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).

The exploratory endpoints were foreseen to evaluate overall accuracy analyses of pathologist reviewed AIM-NASH, per score component, in clinical groups defined by (where available and relevant) trial of origin, time-point, and NASH treatment. Also, the usefulness of the overlays was explored.

Study Results:

Of the finally enrolled cases, less than 4% of the slides had missing final GT score due to various reasons (such as sample, stain, or scan inadequacy); most being for fibrosis (3.2%) and least being for steatosis (1.33%). Evaluations were performed on the available scorings with numbers ranging from 1429 (for AI-assisted fibrosis vs. GT) to 1481 (for IMR steatosis vs GT).

There was an issue based on differential description of the datasets in the protocol (and shown in the briefing document), which could finally be clarified to be an erroneous copying of the protocol numbers. In fact, the final study report describes the finally available cases that were transferred to the Applicant for the use of the study by the trial sponsor adequately. One of the phase 2 pegbelfermin (FALCON 1) studies was not included in the clinical validation study. Overall, for FALCON 2, 154 patients with 284 unique samples were enrolled. For REGENERATE, 470 unique subjects with 694 samples, representative of screened and enrolled populations, and for the Novo Nordisk semaglutide study, 523 unique samples were available, and all were enrolled.

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

Less than 4% of the slides had a missing score from the AI-assisted workflow due to the pathologists unable to score the slide. In addition, there were 7 slides where AIM-NASH was not able to provide a score due to blurry images.

The dataset included represents both, screen-failed as well as enrolled MASH clinical trial patient populations, including study subjects who may have regressed or progressed during a clinical trial, and reflects the MASH 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).

However, there is again underrepresentation of earlier/mild stages of the disease (e.g. for steatosis only 8% with grade 0), only 0.8% with grade 0 lobular inflammation, and 0.9% with stage 0 fibrosis.

Also, lobular inflammation grade 3 is underrepresented with only 2%.

The Applicant was therefore asked to discuss the potential implications and consequences of such an under-representation. In their response, The Applicant referred to the overall representativeness of the included trials (and the population) for current NASH clinical trials with different drug candidates, and different development phases. In this respect, it is admitted that NASH clinical trials – due to the requirements laid down in the guidance documents of both FDA and CHMP/EMA, trials will enrol mainly (or exclusively) patients with NAS>4 and fibrosis stages 2 and above. Therefore, the argument is well taken for the repeated evaluation of histology after treatment, since patients with less severe inflammation, ballooning and fibrosis will not be included in the trials. However, the argument is not fully true for the screening histology evaluations for which an increase of screen-failed cases could have helped for this case.

In their further response, the Applicant has provided additional evaluations for different categories (such as F0&F1 vs. other fibrosis stages, F2&F3 vs other fibrosis stages, F4 vs. other fibrosis stages, NAS ≥4 and ≥1 for each component vs. Other, and NASH resolution). The results are displayed further down below.

Additional tables were presented for the "accuracy" evaluations for percent agreement of different grading and staging (see also below). The Applicant in view of the variability of these results pointed out that the study was not designed to be powered for every score level, and the fact that there is no relevant literature available describing the accuracy and variability of histology evaluations according to the different scores/stages.

Regarding the potential impact, the Applicant also stated that, according to the results, WK results demonstrate non-inferiority for all features and superiority for NAS >=4, as well as NASH resolution, and % agreement demonstrate superiority for AIM-NASH in (positively) identifying F2 and 3 populations, NAS>=4, and NASH resolution across the entire CV dataset.

It is important to note that, for the proposed context of use in both enrolment scoring and evaluation of histologic change for primary endpoint, performance must satisfy both high levels of accuracy and consistency or reproducibility requirements.

It was concluded that the combination of:

1. the accuracy demonstrated overall for steatosis, fibrosis, inflammation, and ballooning, and for the specific clinical trial composite scores comprising a large range of disease activity with varying individual histologic component scores and

2. the superior repeatability/reproducibility of AIM-NASH compared to manual pathology (intra- and inter-)

should result in more accurate, standardized and consistent enrolment and detection of steatosis, ballooning, and/or inflammation grade change or fibrosis stage change for a patient in a trial.

This argumentation, in conjunction with the additional data provided were concluded to be adequate (see also below)

The Applicant also referred to the results of the published case studies (see following chapter).

Similar to the other studies, an "observed cases" evaluation was performed. Since this is the final "pivotal" trial for the validation of the algorithm, the Applicant was asked to address the adequacy of this mode of evaluation.

The Applicant has re-evaluated the "ITT-conform" WKs (replacing missing values, if applicable), using a worst- and best-case imputation for the manual and the PathAI readings, for both the analytical as well as the clinical validation studies. When worst-imputation was used for both methods, both the AV, and the CV showed statistically significant results with superiority of the AIM-NASH tool (which was, however, "borderline" for steatosis in the AV only). As expected, a similar result is achieved when AIM-NASH was evaluated using "best case imputation" and IMR using worst case imputation. Using a "best case imputation" approach for both, the AIM-NASH and the IMR, the results were somewhat influenced by this imputation method, showing superiority of the AIM-NASH tool for hepatocellular ballooning only, but non-inferiority for the other features (being "borderline" of steatosis in the AV). When using the worst-case imputation for AIM-NASH, and best-case imputation for IMR, a more "diverse" picture emerged with statistical superiority of the AIM-NASH vs. IMR only achieved for ballooning (all three analyses presented), and for lobular inflammation in the CV. Non-inferiority in this evaluation, however, was concluded in all evaluations, except for steatosis in the AV, which was "borderline" significant for non-inferiority only. Considering the aims of the trials, it was indeed demonstrated that the conclusion on non-inferiority (for the comparison of WKs) was remarkably robust against the influence of missing values. Based on these additional data, the presented primary evaluation was accepted.

The results of the primary evaluation are shown in the following table.

Table 22: 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,	<0.0001	0.42

	IMR vs GT	1453	0.645 (0.622, 0.665)	0.039)		

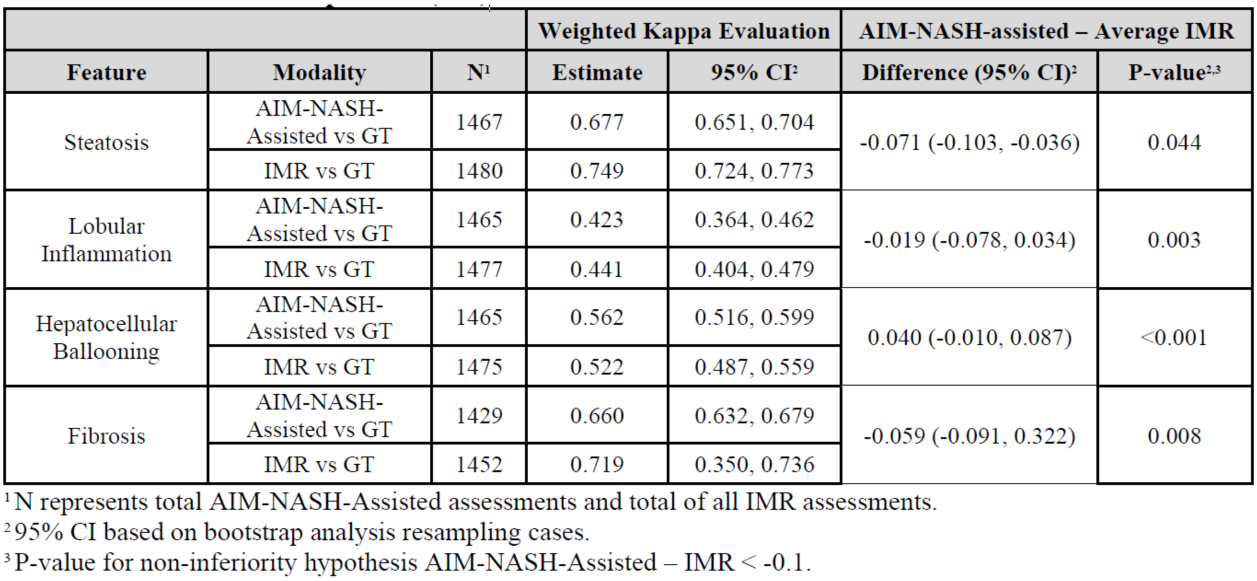
As obvious from the table, the AIM-NASH evaluation performed superior to the IMR evaluation regarding lobular inflammation and hepatocellular ballooning, and non-inferior with regard to the features steatosis and fibrosis. There was literally no "net gain" for reduced variability (individual manual read compared to GT vs. AI-assist compared to GT) for steatosis and fibrosis, since the WKs were almost the same. The difference for the WKs in lobular inflammation and ballooning, however, seem to be clinically relevant.

The trial results (and the reported WKs) raised several questions on the "external validity" or contextualisation of the study. Two issues were further discussed. The first related to the achieved "absolute" values for the WKs in comparison to those reported in the literature, and the second related to the "validity" of the comparisons made, on the background that the comparison to the IMR panel may not be the most relevant analysis, but the performance in comparison to the GT itself may be much more relevant.

1. Comparison to statistical "ground truth"

To address the concern, the Applicant conducted a further analysis, comparing the AIM-NASH assisted reads against a "statistical (median) GT consensus using mode" with the median IMR consensus against the median GT. Both the GT and IMR consensus modes were defined with a method that has recently been used in clinical MASH trials, where the final read for a score component is the mode (2 out of 3 pathologists' agreement on a score component), and the median is used in the case that all three scores provided are different. Hence a different mode was applied to the GT and the IMR scores as defined in the primary evaluation to generate 2 different statistical consensus panel scores per case, and this evaluation compared to the AIM-NASH based evaluation. This "explorative" evaluation is shown in the following table.

Table 23: Primary endpoint results for each NASH histologic component/explorative evaluation using "Median Panel Comparison"



			Weighted Kappa Evaluation		AIM-NASH-assisted – Average IMR	
Feature	Modality	N[1]	Estimate	95% CI[2]	Difference (95% CI)[2]	P-value[2,3]
Steatosis	AIM-NASH-Assisted vs GT	1467	0.677	0.651, 0.704	-0.071 (-0.103, -0.036)	0.044
	IMR vs GT	1480	0.749	0.724, 0.773		
Lobular Inflammation	AIM-NASH-Assisted vs GT	1465	0.423	0.364, 0.462	-0.019 (-0.078, 0.034)	0.003
	IMR vs GT	1477	0.441	0.404, 0.479		
Hepatocellular Ballooning	AIM-NASH-Assisted vs GT	1465	0.562	0.516, 0.599	0.040 (-0.010, 0.087)	<0.001
	IMR vs GT	1475	0.522	0.487, 0.559		
Fibrosis	AIM-NASH-Assisted vs GT	1429	0.660	0.632, 0.679	-0.059 (-0.091, 0.322)	0.008
	IMR vs GT	1452	0.719	0.350, 0.736		

[1] N represents total AIM-NASH-Assisted assessments and total of all IMR assessments.
[2] 95% CI based on bootstrap analysis resampling cases.
[3] P-value for non-inferiority hypothesis AIM-NASH-Assisted – IMR < -0.1.

In this mode/median analysis, the primary endpoint of non-inferiority (WK within ±0.1) was met for all histologic components for AIM-NASH-assisted reads compared to statistical consensus reads. Although the AIM-NASH performed numerically somewhat inferior to the "Median Consensus Mode", this explorative analysis overall indicated that AIM-NASH agreed with the statistical GT consensus in a non-inferior manner as compared to the agreement between two different statistical consensus groups (median IMR and median GT) and was taken as overall reassuring regarding the robustness of the tool.

2. External comparison/contextualisation

The WKs reported in the literature (Shrout-Fleiss method) were in the range of 0.75-0.8 (studies referenced Kleiner 2005 and 2019, Sanyal 2021) for fibrosis, while the AIM-NASH and IMR vs. GT evaluation ended up with an overall 0.65 WK value Similarly, the literature reported values for steatosis are between 0.77 and 0.89 in the three studies, while the current study ended up with a WK of 0.67. Lobular inflammation has been reported with WKs around 0.45 (by the two Kleiner studies) up to 0.60 (Sanyal), while only 0.42 were achieved in the current trial with AIM-NASH against GT evaluation. It was also noted that in the literature reports, different "types" of WK-values were reported (especially as seen in Sanyal 2021), and clarification was therefore requested whether an adequate comparison was provided in the first place.

The comparison to the literature as initially presented by the Applicant is shown in the following table:

Table 24: 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.*

The reference by Sanyal (2021) on which the conclusion on different types of WKs, with a similar panel to panel comparison was based, was reporting the following:

Table 25: Inter-reader Agreement for NASH Histologic Features from the NASH-CRN

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

In their response, first of all, the Applicant referred to the fact that the WK values reported in the analytical and clinical validation studies were linearly WKs (Cicchetti-Allison), and that the same linearly WKs were also reported in published papers (Davison et al. 2020), (Kleiner et al. 2005) and (Kleiner et al. 2019) and are therefore directly comparable as far as agreement methodology is concerned. In the reference Sanyal et al. 2021, the WKs from (Kleiner et al. 2005) and (Kleiner et al. 2019) are incorrectly labelled as quadratic WKs (Shrout-Fleiss), for which the Applicant has provided evidence with directly contacting the authors. It is therefore concluded that the values reported in the panel evaluation in the Sanyal 2021 study (with a modified "median consensus method"), comparing two panel with evaluation with the same method, were not exceptionally better than what has historically been reported in the NAS-CRN studies (Kleiner 2005 and Kleiner 2019). Only some improvement of the WK value as compared to the Kleiner studies can be seen for steatosis (for which variability is, however, low anyway). Still, the NAS-CRN reported WKs, as well as those reported in the Sanyal study ("Panel A vs B") appear numerically partly relevantly distinct from what was achieved in the clinical validation study as presented. While for lobular inflammation and ballooning the WKs appear to be in the same range, there seems to be a relevant difference to the disadvantage of the AIM-NASH evaluated WKs for fibrosis and steatosis as compared to the Kleiner study.

The Applicant has additionally provided argumentation regarding the differences of the studies, referring to sample sizes, and – more importantly – the comparisons made. While the Davison paper compares pairwise WKs (out of three pathologists), the Kleiner studies have reported average pairwise WKs across a whole group of up to nine pathologists. Contrary to this, the Sanyal study reported WKs across two different panels of three pathologists (with WKs between individuals). Also, the Applicant refers to the potential for higher variability considering the variety of data from multiple studies included in the AV and CV studies. The Applicant also takes the kappas from the panel-to-panel comparisons in the Sanyal paper and the exploratory median consensus analysis as an argument to demonstrate that the gold standard panel workflow still experiences intra- and inter- variability, like that seen with expert individual CRN pathologists. The comparisons in the analytical and clinical validation studies therefore can stand on their own, and the "external comparisons" are only suitable to give some orientation of the level of variance that is achievable with the AIM-NASH tool.

Regarding differences between trials, it can also be referred to the differences seen e.g. in inflammation and fibrosis score WKs e.g. for the FALCON 1 and 2 trials as compared to the REGENERATE trial as seen in the analytical validation (see above).

This argumentation was considered adequate, and there was general acknowledgement that between study comparison, especially when done on different datasets and using different methodology must generally be handled with caution. In this regard the consistency of the results from the analytical as well as the clinical validation studies, and the robustness of the results even when compared to a different method of evaluation than the IMR were considered appropriate to adequately support the validity of the AIM-NASH methodology.

The obvious differences in the IMR evaluations (vs. GT) seen in the AV and CV for inflammation (WKs in CV validation 0.297 and 0.402 in the AV) was also discussed with the Applicant. The Applicant again referred to differences between the trials used (e.g. CV had a significantly larger sample size, inclusion of different (additional) drug candidates in the CV), and hence the dataset composition and the different case numbers. The Applicant concluded that the representativeness of the CV dataset was clearly higher than the AV dataset and results would therefore be considered to be more robust. The Applicant also referred to the fact that the rates of overall positive agreement in the AV and CV studies for the IMR vs GT did not change for fibrosis (AV 56.4% vs. CV 54.1%), while the AIM-NASH vs GT agreement rates increased from 57% to 63.5%, and hence it can be excluded that the superiority of the AIM-NASH is only due to a poor performance of the IMR evaluations. This argumentation was considered acceptable, and the issue was considered sufficiently resolved.

Secondary endpoint:

The secondary endpoint for the "categorical evaluation" of the NASH features (F0&F1 vs. other; F4 vs. other, NAS≥4 vs. NAS <4), the following results were achieved:

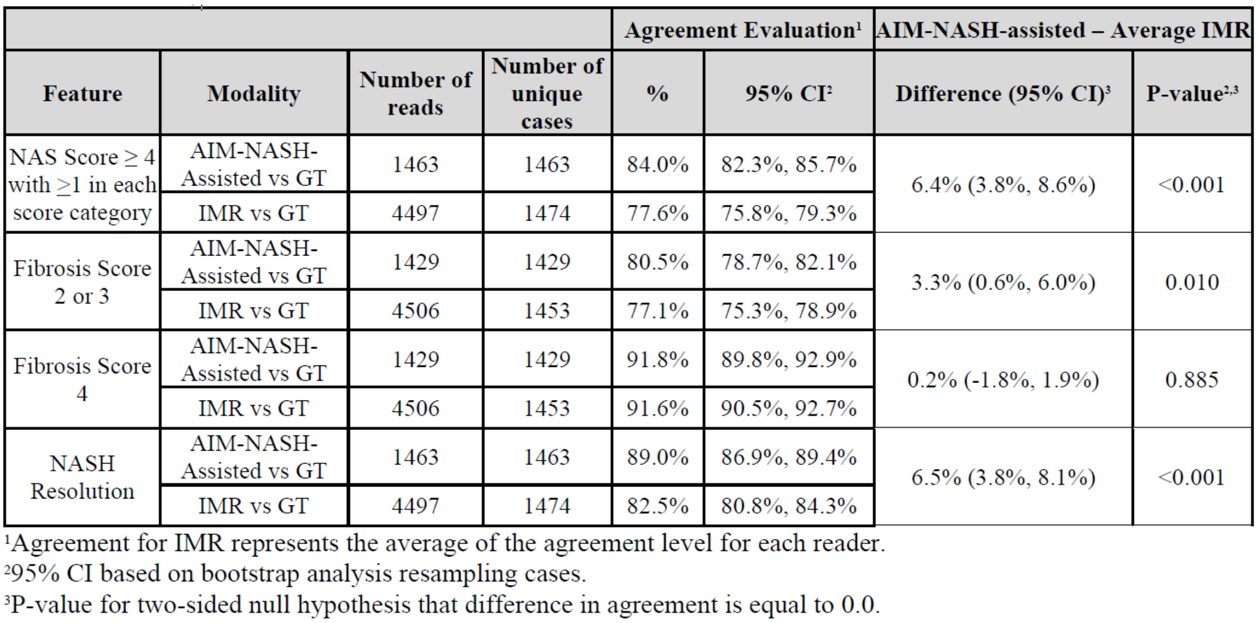
Table 26: WKs for aggregate NASH component scores

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)	

As obvious from the table, although the description of the Applicant that the WKs were grossly similar with overlapping confidence intervals, there is a statistically significant superiority regarding the "NAS ≥4 vs. <4" category, however, it is also obvious that the aim of achieving a lower bound of the 97.5% CI of 0.1 has not been achieved for the F0&F1 category.

The Applicant was also asked to provide a more elaborate evaluation on this "categorical evaluation", and provided the following additional data, displaying rates of positive and negative agreement, as well as overall agreement.

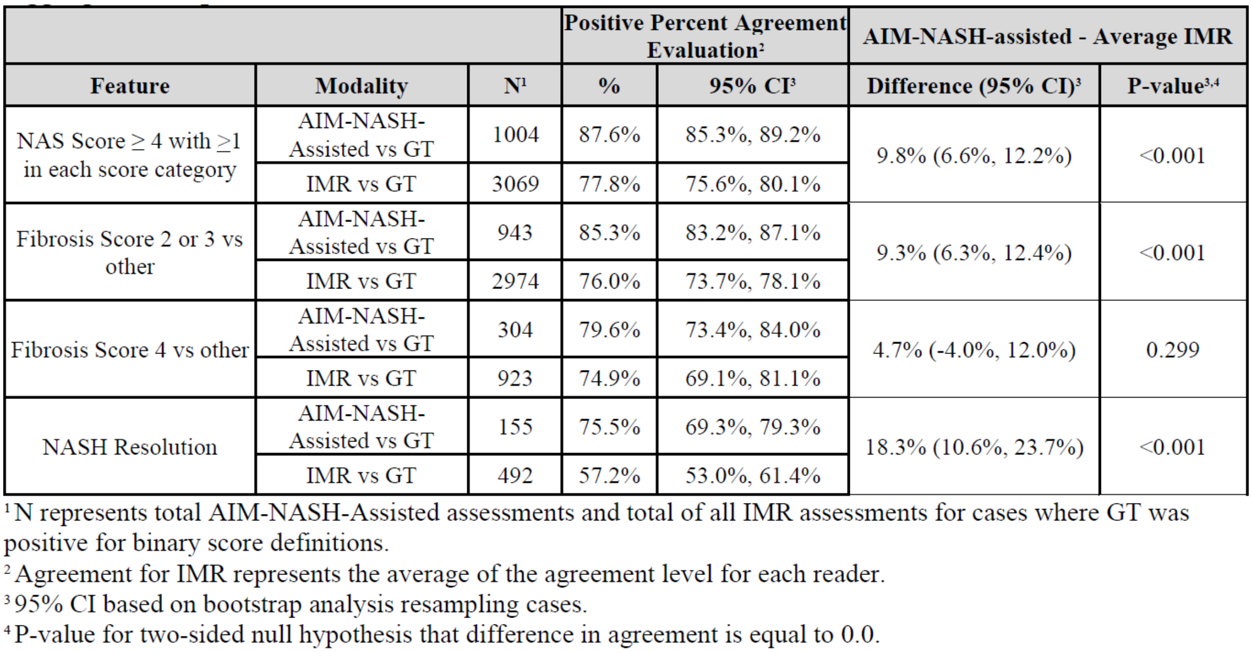
Table 27: Overall % Agreement with Ground Truth for Aggregate Component Scores (CV AIM NASH-assisted)



				Agreement Evaluation[1]		AIM-NASH-assisted – Average IMR	
Feature	Modality	Number of reads	Number of unique cases	%	95% CI[2]	Difference (95% CI)[3]	P-value[2,3]
NAS Score ≥ 4 with ≥1 in each score category	AIM-NASH-Assisted vs GT	1463	1463	84.0%	82.3%, 85.7%	6.4% (3.8%, 8.6%)	<0.001
	IMR vs GT	4497	1474	77.6%	75.8%, 79.3%		
Fibrosis Score 2 or 3	AIM-NASH-Assisted vs GT	1429	1429	80.5%	78.7%, 82.1%	3.3% (0.6%, 6.0%)	0.010
	IMR vs GT	4506	1453	77.1%	75.3%, 78.9%		
Fibrosis Score 4	AIM-NASH-Assisted vs GT	1429	1429	91.8%	89.8%, 92.9%	0.2% (-1.8%, 1.9%)	0.885
	IMR vs GT	4506	1453	91.6%	90.5%, 92.7%		
NASH Resolution	AIM-NASH-Assisted vs GT	1463	1463	89.0%	86.9%, 89.4%	6.5% (3.8%, 8.1%)	<0.001
	IMR vs GT	4497	1474	82.5%	80.8%, 84.3%		

[1] Agreement for IMR represents the average of the agreement level for each reader.
[2] 95% CI based on bootstrap analysis resampling cases.
[3] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

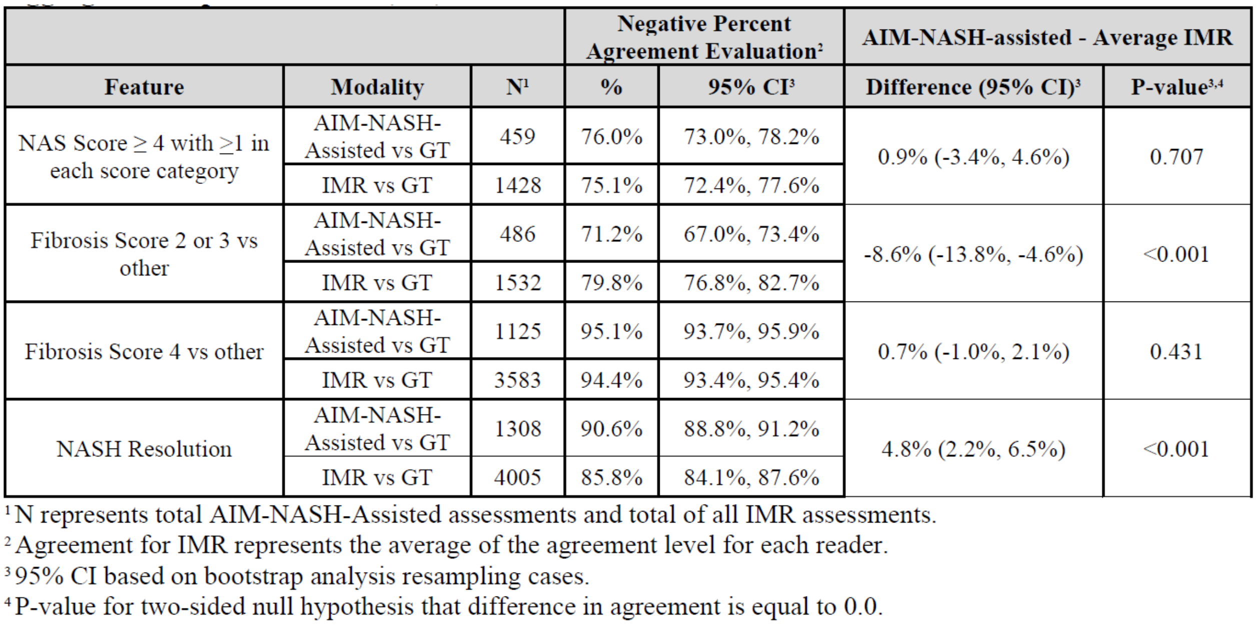
Table 28: Positive % Agreement with Ground Truth for Aggregate Component Scores (CV AIMNASH-assisted)



			Positive Percent Agreement Evaluation[2]		AIM-NASH-assisted - Average IMR	
Feature	Modality	N[1]	%	95% CI[3]	Difference (95% CI)[3]	P-value[3,4]
NAS Score ≥ 4 with ≥1 in each score category	AIM-NASH-Assisted vs GT	1004	87.6%	85.3%, 89.2%	9.8% (6.6%, 12.2%)	<0.001
	IMR vs GT	3069	77.8%	75.6%, 80.1%		
Fibrosis Score 2 or 3 vs other	AIM-NASH-Assisted vs GT	943	85.3%	83.2%, 87.1%	9.3% (6.3%, 12.4%)	<0.001
	IMR vs GT	2974	76.0%	73.7%, 78.1%		
Fibrosis Score 4 vs other	AIM-NASH-Assisted vs GT	304	79.6%	73.4%, 84.0%	4.7% (-4.0%, 12.0%)	0.299
	IMR vs GT	923	74.9%	69.1%, 81.1%		
NASH Resolution	AIM-NASH-Assisted vs GT	155	75.5%	69.3%, 79.3%	18.3% (10.6%, 23.7%)	<0.001
	IMR vs GT	492	57.2%	53.0%, 61.4%		

[1] N represents total AIM-NASH-Assisted assessments and total of all IMR assessments for cases where GT was positive for binary score definitions.
[2] Agreement for IMR represents the average of the agreement level for each reader.
[3] 95% CI based on bootstrap analysis resampling cases.
[4] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 29: Negative % Agreement with Ground Truth for Aggregate Component Scores (CV AIMNASH-assisted)



			Negative Percent Agreement Evaluation[2]		AIM-NASH-assisted - Average IMR	
Feature	Modality	N[1]	%	95% CI[3]	Difference (95% CI)[3]	P-value[3,4]
NAS Score ≥ 4 with ≥1 in each score category	AIM-NASH-Assisted vs GT	459	76.0%	73.0%, 78.2%	0.9% (-3.4%, 4.6%)	0.707
	IMR vs GT	1428	75.1%	72.4%, 77.6%		
Fibrosis Score 2 or 3 vs other	AIM-NASH-Assisted vs GT	486	71.2%	67.0%, 73.4%	-8.6% (-13.8%, -4.6%)	<0.001
	IMR vs GT	1532	79.8%	76.8%, 82.7%		
Fibrosis Score 4 vs other	AIM-NASH-Assisted vs GT	1125	95.1%	93.7%, 95.9%	0.7% (-1.0%, 2.1%)	0.431
	IMR vs GT	3583	94.4%	93.4%, 95.4%		
NASH Resolution	AIM-NASH-Assisted vs GT	1308	90.6%	88.8%, 91.2%	4.8% (2.2%, 6.5%)	<0.001
	IMR vs GT	4005	85.8%	84.1%, 87.6%		

[1] N represents total AIM-NASH-Assisted assessments and total of all IMR assessments.
[2] Agreement for IMR represents the average of the agreement level for each reader.
[3] 95% CI based on bootstrap analysis resampling cases.
[4] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

The Applicant concludes on these data, that agreement results demonstrate either equivalency or superiority for all composite scores evaluated. For NAS >=4, F2/3, as well as NASH resolution, in the % agreement tables demonstrate superiority for AIM-NASH algorithm and/or AIM-NASH assisted reads in identifying F2, 3 populations, NAS>=4, and NASH resolution across the entire CV dataset. It is important to note that, for this proposed context of use, performance must satisfy both high levels of accuracy and consistency or reproducibility requirements.

The Applicant further concludes that the combination of:

1. the accuracy (OPA, PPA, and NPA) demonstrated overall for steatosis, fibrosis, inflammation, and ballooning, and for the specific clinical trial composite scores comprising a large range of disease activity with varying individual histologic component scores and

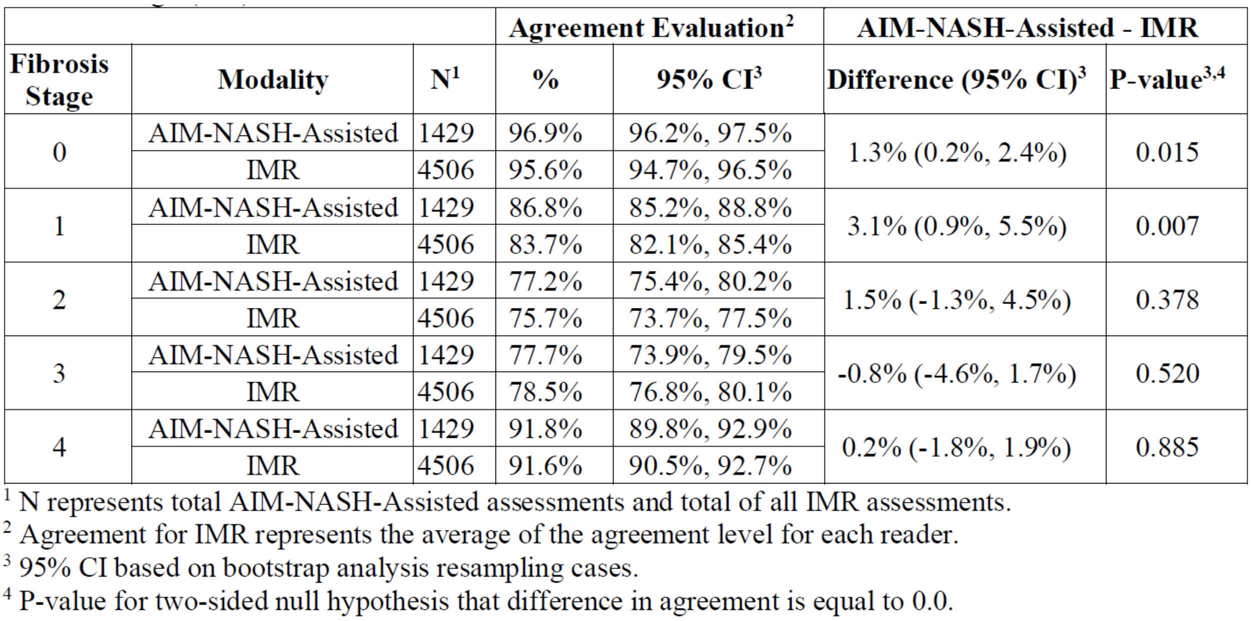
2. the superior repeatability/reproducibility of AIM-NASH compared to manual pathology (intra- and inter-)

should result in more accurate, standardized and consistent enrolment and detection of steatosis grade change or fibrosis stage change for a patient in a trial

The tables indeed regularly show a more "accurate" designation of the NAS based inclusion criteria (usually used in clinical trials in MASH), as well as for the designation of one part of the "interim" primary endpoint, the NASH resolution. Fibrosis on the other hand, appears to be "stable" in the sense that accuracy achieves the same level for AIM-NASH as for average IMR evaluations, which is likewise reassuring (fibrosis stage is used as inclusion criterion, as well as intermediate endpoint). In addition, the similar accuracy of the fibrosis score 4 vs other points to the possibility that the tool is also adequate for the composite "hard" endpoint in trials with non-cirrhotic NASH (=manifestation of cirrhosis), although the validation exercise was not designed to evaluate this part in confirmatory manner, and the use for the final endpoint evaluation of cirrhosis will not be part of the context of use statement. The only drawback in the analyses presented was the inferior performance for the NPA for fibrosis scores 2-3 which was further discussed with regard to its potential impact on the interim endpoint evaluation in MASH clinical trials.

The descriptive analysis of the agreement rates according to fibrosis stage showed the following results:

Table 30: Overall % Agreement with Ground Truth by Level of Fibrosis Stage (CV AIMNASH-Assisted vs IMR)

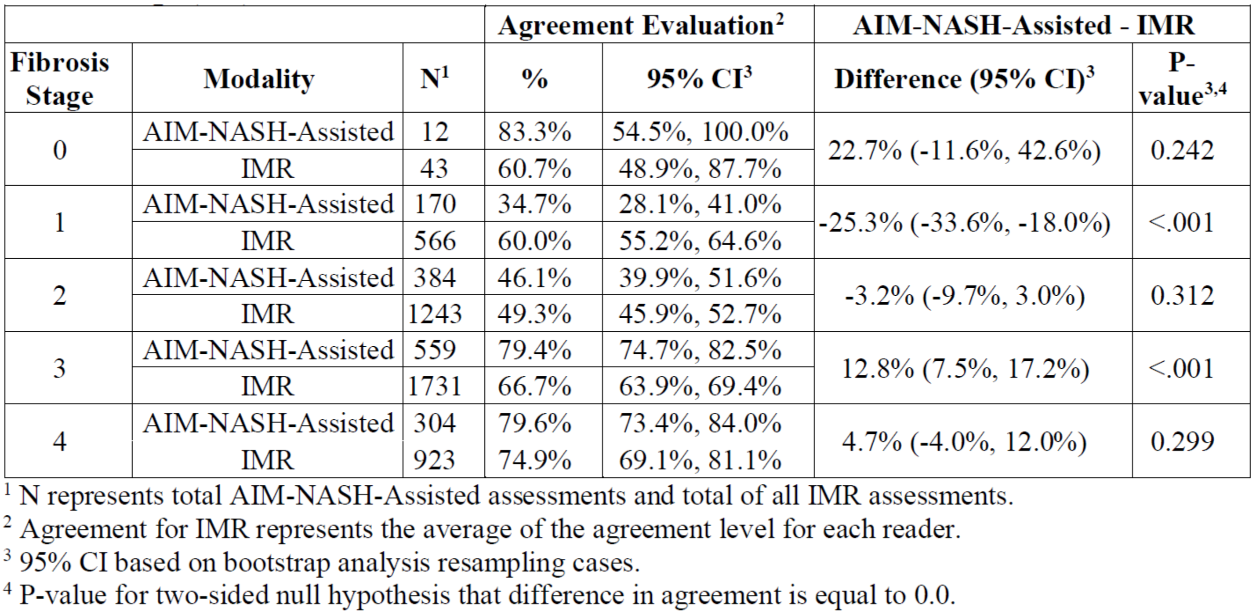


			Agreement Evaluation[2]		AIM-NASH-Assisted - IMR	
Fibrosis Stage	Modality	N[1]	%	95% CI[3]	Difference (95% CI)[3]	P-value[3,4]
0	AIM-NASH-Assisted	1429	96.9%	96.2%, 97.5%	1.3% (0.2%, 2.4%)	0.015
	IMR	4506	95.6%	94.7%, 96.5%		
1	AIM-NASH-Assisted	1429	86.8%	85.2%, 88.8%	3.1% (0.9%, 5.5%)	0.007
	IMR	4506	83.7%	82.1%, 85.4%		
2	AIM-NASH-Assisted	1429	77.2%	75.4%, 80.2%	1.5% (-1.3%, 4.5%)	0.378
	IMR	4506	75.7%	73.7%, 77.5%		
3	AIM-NASH-Assisted	1429	77.7%	73.9%, 79.5%	-0.8% (-4.6%, 1.7%)	0.520
	IMR	4506	78.5%	76.8%, 80.1%		
4	AIM-NASH-Assisted	1429	91.8%	89.8%, 92.9%	0.2% (-1.8%, 1.9%)	0.885
	IMR	4506	91.6%	90.5%, 92.7%		

[1] N represents total AIM-NASH-Assisted assessments and total of all IMR assessments.
[2] Agreement for IMR represents the average of the agreement level for each reader.
[3] 95% CI based on bootstrap analysis resampling cases.
[4] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

* Note that the N's in the above table are for overall number of comparisons (e.g., total agreement on F0 vs. Other, F1 vs. other, etc.), not representative of the number of slides for each score level.

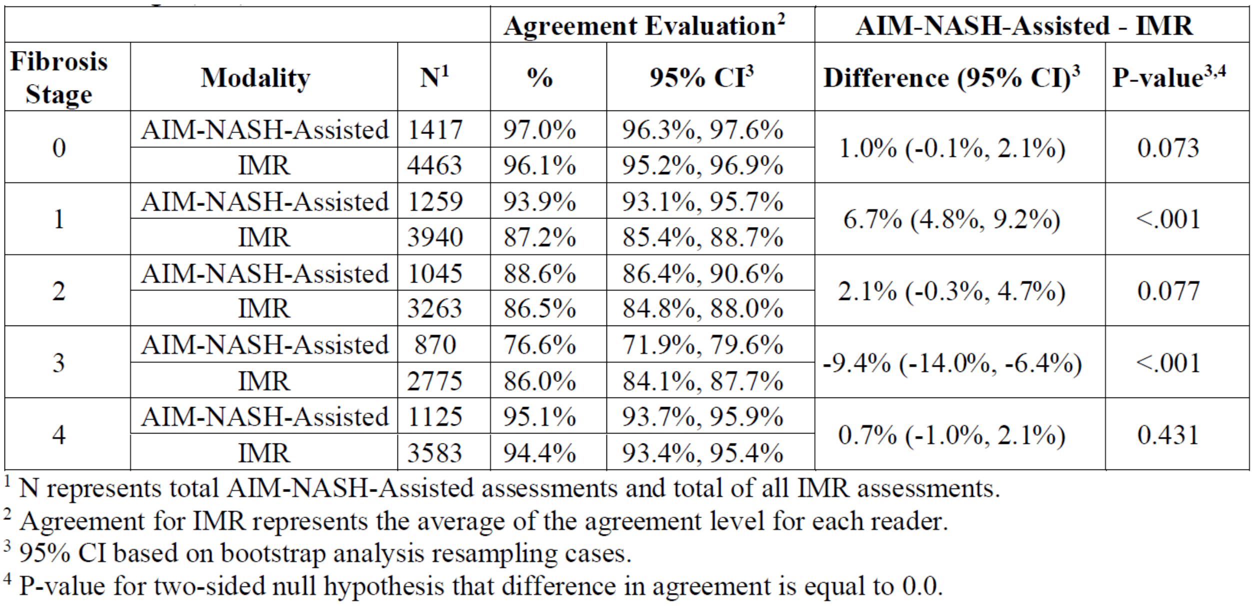
Table 31: Positive % Agreement with Ground Truth by Level of Fibrosis Stage (CV AIMNASH-Assisted vs IMR)



			Agreement Evaluation[2]		AIM-NASH-Assisted - IMR	
Fibrosis Stage	Modality	N[1]	%	95% CI[3]	Difference (95% CI)[3]	P-value[3,4]
0	AIM-NASH-Assisted	12	83.3%	54.5%, 100.0%	22.7% (-11.6%, 42.6%)	0.242
	IMR	43	60.7%	48.9%, 87.7%		
1	AIM-NASH-Assisted	170	34.7%	28.1%, 41.0%	-25.3% (-33.6%, -18.0%)	<.001
	IMR	566	60.0%	55.2%, 64.6%		
2	AIM-NASH-Assisted	384	46.1%	39.9%, 51.6%	-3.2% (-9.7%, 3.0%)	0.312
	IMR	1243	49.3%	45.9%, 52.7%		
3	AIM-NASH-Assisted	559	79.4%	74.7%, 82.5%	12.8% (7.5%, 17.2%)	<.001
	IMR	1731	66.7%	63.9%, 69.4%		
4	AIM-NASH-Assisted	304	79.6%	73.4%, 84.0%	4.7% (-4.0%, 12.0%)	0.299
	IMR	923	74.9%	69.1%, 81.1%		

[1] N represents total AIM-NASH-Assisted assessments and total of all IMR assessments.
[2] Agreement for IMR represents the average of the agreement level for each reader.
[3] 95% CI based on bootstrap analysis resampling cases.
[4] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

Table 32: Negative % Agreement with Ground Truth by Level of Fibrosis Stage (CV AIMNASH-Assisted vs IMR)



			Agreement Evaluation[2]		AIM-NASH-Assisted - IMR	
Fibrosis Stage	Modality	N[1]	%	95% CI[3]	Difference (95% CI)[3]	P-value[3,4]
0	AIM-NASH-Assisted	1417	97.0%	96.3%, 97.6%	1.0% (-0.1%, 2.1%)	0.073
	IMR	4463	96.1%	95.2%, 96.9%		
1	AIM-NASH-Assisted	1259	93.9%	93.1%, 95.7%	6.7% (4.8%, 9.2%)	<.001
	IMR	3940	87.2%	85.4%, 88.7%		
2	AIM-NASH-Assisted	1045	88.6%	86.4%, 90.6%	2.1% (-0.3%, 4.7%)	0.077
	IMR	3263	86.5%	84.8%, 88.0%		
3	AIM-NASH-Assisted	870	76.6%	71.9%, 79.6%	-9.4% (-14.0%, -6.4%)	<.001
	IMR	2775	86.0%	84.1%, 87.7%		
4	AIM-NASH-Assisted	1125	95.1%	93.7%, 95.9%	0.7% (-1.0%, 2.1%)	0.431
	IMR	3583	94.4%	93.4%, 95.4%		

[1] N represents total AIM-NASH-Assisted assessments and total of all IMR assessments.
[2] Agreement for IMR represents the average of the agreement level for each reader.
[3] 95% CI based on bootstrap analysis resampling cases.
[4] P-value for two-sided null hypothesis that difference in agreement is equal to 0.0.

The Applicant has provided an elaborate response on this question (additional analyses were also presented as compared to those above), which not only displayed the data requested (agreement rates for the different fibrosis stages), but also an analysis of the potential impact of the observed higher variability in the early fibrosis stages. It has been shown that overall agreement rates are very similar to the IMR assessment, even for the "early" fibrosis stages, but that AIM-NASH performs "worse" in identifying fibrosis stage 1, and "worse" in excluding fibrosis stage 3. Whether this is only a play of chance, or whether this is a "systemic" result remains debatable.

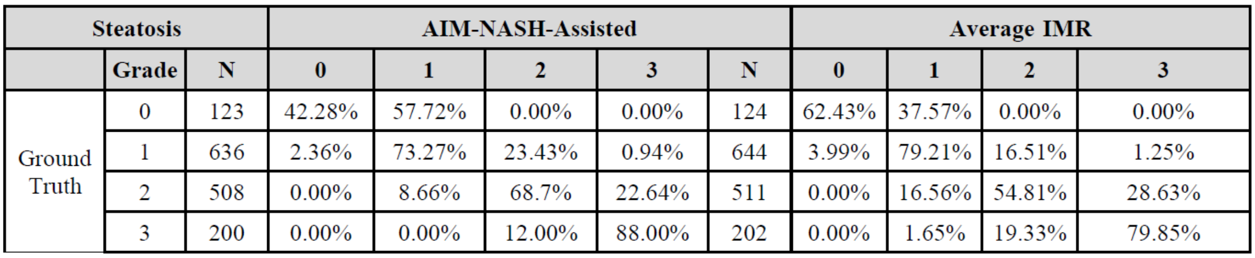
One of the consequences of the results was also again debated, with reference to the tables submitted with the response to the first List of Issues which showed that the overall F2/F3 PPA was indeed higher for AIM-NASH as compared to the IMR reads (difference +9.3%), but the NPA was lower (-8.6%). Again, the identification of F2/3 was better, but exclusion of F2/3 worse than obtained by IMR reads. It is therefore concluded and agreed with the Applicant that more patients are identified with F2/F3 fibrosis stages with AIM-NASH as compared to an IMR evaluation.

The potential impact of these performance characteristics where then analysed by looking into the response rates achieved in the FALCON 2 and REGENERATE trials analysed by IMR or AIM-NASH. This showed that the agreement with the results of the two trials (taken as ground truth) was not relevantly different, with a small percentage indicating somewhat lower response rates with the AIM-NASH analysis. However, both analyses had a "tendency" for overall increased response rates, as compared to the ground truth, which was higher in the FALCON 2 trial.

Whether this is a "systematic" property of the tool, however, can also be evaluated based on the presented "case studies" included in the primary submission, as well as the additional study meanwhile submitted (see below). Looking at these results, however, there is no tendency to have overall higher responder rates regarding fibrosis improvement (in those studies for which this endpoint was evaluated). The generation of further "cases" may shed additional light on this open issue.

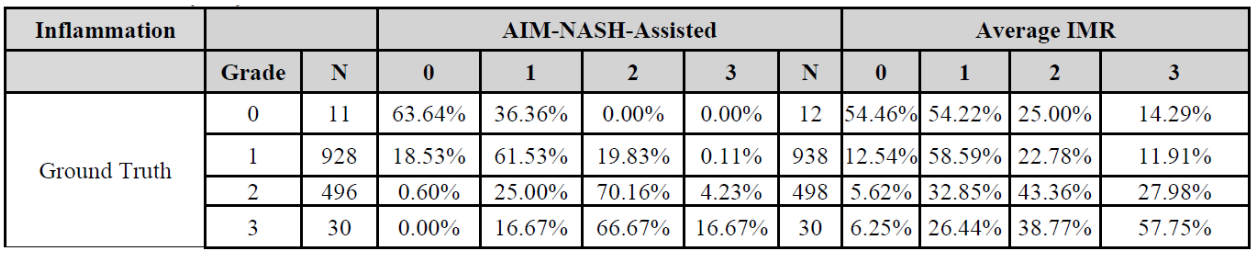
As mentioned above, in the context of the low representation of certain subgroups of patients, the Applicant has provided exploratory data within additional tables for "% agreement" in different histology categories for all four histological features evaluated. Please note, the number of samples in some groups is low, and the data, therefore, should be interpreted with caution.

Table 33: AIM-NASH-Assisted and IMR Agreement with Ground Truth per Score for Steatosis



Steatosis			AIM-NASH-Assisted					Average IMR			
	Grade	N	0	1	2	3	N	0	1	2	3
Ground Truth	0	123	42.28%	57.72%	0.00%	0.00%	124	62.43%	37.57%	0.00%	0.00%
	1	636	2.36%	73.27%	23.43%	0.94%	644	3.99%	79.21%	16.51%	1.25%
	2	508	0.00%	8.66%	68.7%	22.64%	511	0.00%	16.56%	54.81%	28.63%
	3	200	0.00%	0.00%	12.00%	88.00%	202	0.00%	1.65%	19.33%	79.85%

Table 34: AIM-NASH-Assisted and IMR Agreement with Ground Truth per Score for Lobular Inflammation



Inflammation			AIM-NASH-Assisted					Average IMR			
	Grade	N	0	1	2	3	N	0	1	2	3
Ground Truth	0	11	63.64%	36.36%	0.00%	0.00%	12	54.46%	54.22%	25.00%	14.29%
	1	928	18.53%	61.53%	19.83%	0.11%	938	12.54%	58.59%	22.78%	11.91%
	2	496	0.60%	25.00%	70.16%	4.23%	498	5.62%	32.85%	43.36%	27.98%
	3	30	0.00%	16.67%	66.67%	16.67%	30	6.25%	26.44%	38.77%	57.75%

Table 35: AIM-NASH-Assisted and IMR Agreement with Ground Truth per Score for Hepatocellular Ballooning

Ballooning			AIM-NASH-Assisted				Average IMR		
	Grade	N	0	1	2	N	0	1	2
Ground Truth	0	161	75.78%	23.60%	0.62%	164	62.70%	39.03%	5.05%
	1	689	16.55%	57.62%	25.83%	694	24.95%	56.50%	18.55%
	2	615	1.95%	18.54%	79.51%	618	5.24%	35.46%	59.96%

Table 36: AIM-NASH-Assisted and IMR Agreement with Ground Truth per Score for Fibrosis

Fibrosis			AIM-NASH-Assisted						Average IMR				
	Stage	N	0	1	2	3	4	N	0	1	2	3	4
Ground Truth	0	12	83.33%	8.33%	8.33%	0.00%	0.00%	13	60.65%	37.35%	26.67%	0.00%	0.00%
	1	170	20.00%	34.71%	36.47%	8.82%	0.00%	175	18.96%	60.01%	18.19%	4.28%	1.28%
	2	384	2.08%	18.23%	46.09%	33.33%	0.26%	396	4.78%	31.15%	49.3%	15.58%	1.52%
	3	559	0.00%	1.07%	9.84%	79.43%	9.66%	565	0.44%	4.02%	18.67%	66.62%	12.6%
	4	304	0.00%	0.00%	0.33%	20.07%	79.61%	304	0.00%	0.00%	6.67%	24.23%	74.94%

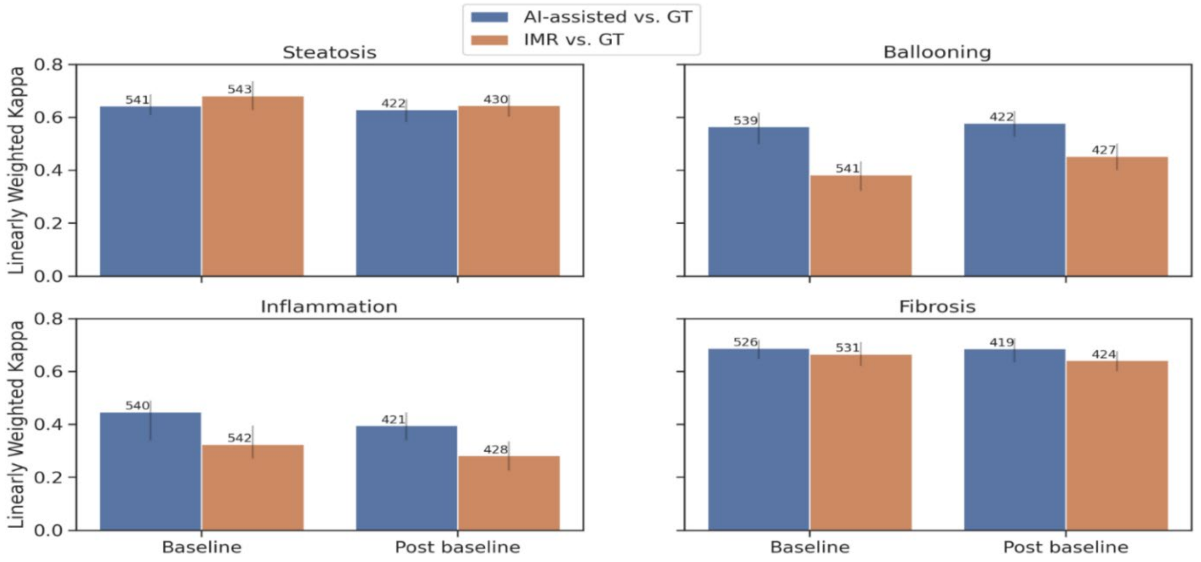
For the assessment, see above.

Exploratory endpoints:

Regarding the time-point evaluation (baseline vs. post-baseline), this 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.

The difference in WKs was usually similar and in agreement with the overall evaluation documenting superiority for lobular inflammation and hepatocellular ballooning for baseline and post-baseline evaluations alike. Also, the keeping with the non-inferiority boundary of 0.1 was fulfilled for steatosis and fibrosis for both time points. The following figure shows the results:

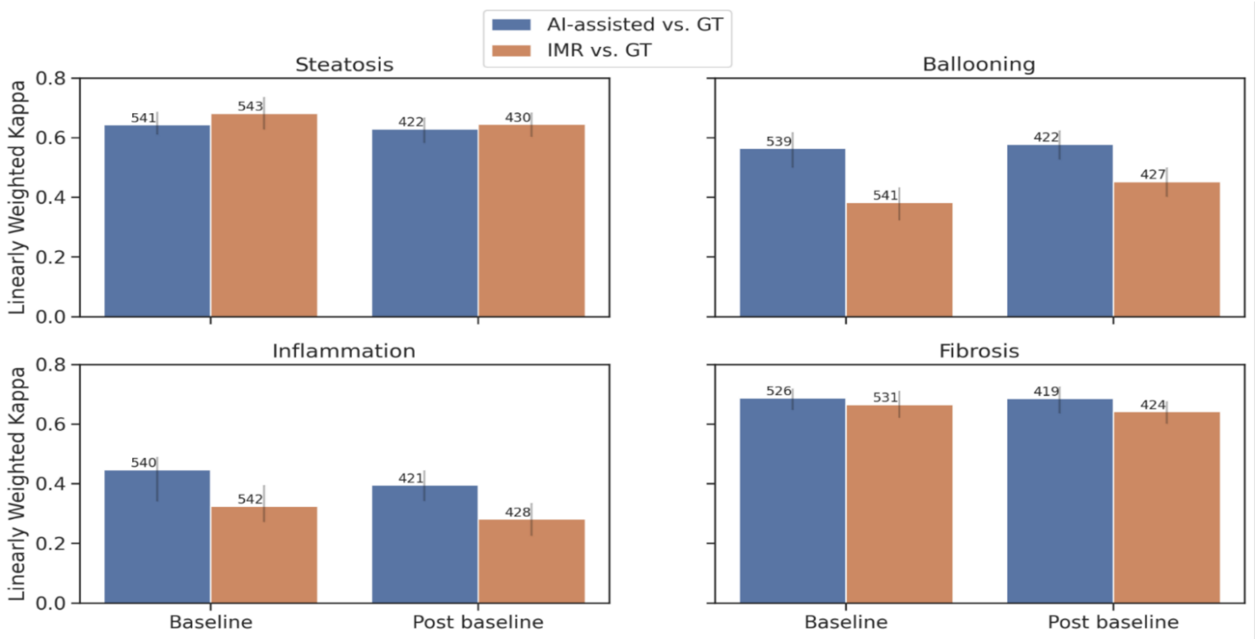
Figure 6: WKs for NASH components per time point for trials with available timepoint data (FALCON 2 and REGENERATE)



For the evaluation by Sponsor, also accordance with the overall results was demonstrated for all components of the evaluation, however, statistically significant superiority regarding lobular inflammation and hepatocellular ballooning was not achieved for all three trials (not in the BMS trial which was the trial with lowest numbers). A similar effect (obviously owing to the lower numbers of the BMS study) could be seen for the non-inferiority evaluation regarding steatosis and fibrosis.

For this phenomenon, the Applicant was asked to discuss the reasons and referred to the fact that in the BMS trial, the number of samples was lower than in the other two trials, and therefore less adequately powered. Additionally, for the BMS trial, the WKs for AIM-NASH-assisted reads are comparable to the WKs of the IMRs, indicating the performance is equivalent to a manual reader. This response was considered overall acceptable. The following figure shows the WK-based comparisons according to trial:

Figure 7: WKs for NASH components per sponsor



Regarding the evaluation according to score level, the overall results are also reflected in the subscores: 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.

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, were considered for this analysis. For all NASH components, between 40% and 50% of the time the pathologists indicated that they were neutral (utility score of 3) to the helpfulness of the overlay. For steatosis and fibrosis, the pathologists found the AIM-NASH overlays useful or very useful approximately 35% of the time, for lobular inflammation 38% of the time and for hepatocellular ballooning 44% of the time. Rarely (between 2% and 5% of the time), the pathologists indicated that the AIM-NASH overlays were not at all useful. It's important to note that the pathologists may toggle the overlays off in the rare instances where they do not find them to be helpful.

Post-hoc exploratory endpoints:

As additional "categorical evaluation" WK comparisons were made for "F2&F3" vs other, NAS≥ 4 with >1 in each score category vs other, and NASH resolution. Since the categories are relevant regarding in- and exclusion criteria (the first 2) and intermediate endpoint evaluation (the third) as currently practised for most MASH trials, and also included in the CHMP Reflection paper, these are displayed in the following table:

Table 37: 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 ≥ 4 with ≥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)	

This evaluation shows that the categories "NAS≥4 with at least 1 in each score category" and NASH resolution performed statistically significantly better in the AIM-NASH evaluation compared to the IMR evaluation (vs GT), and that the F2/3 category was keeping with the non-inferiority requirement of 0.1.

A subset of evaluations was evaluated for cases where the same pathologist reviewed the same slide (the cases verified between 86 and 216 slides per component and were thus including only a minority of evaluations). WKs were computed for this evaluation and showed that there was almost no difference for steatosis and fibrosis, but that lobular inflammation achieved clearly higher WKs, thus reflecting the overall results.

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, 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

Disagreement with AIM-NASH:

1-stage disagreement was relatively frequent with 15%-22% of cases having a disagreement in the four categories; however, 2-stage disagreement was very infrequent with 0.27% (steatosis) - 1.83% (fibrosis).

AIM-NASH accuracy alone

A further (similar evaluation as in the AV) was conducted comparing the WKs of AIM-NASH alone (without input from histopathologist) vs GT compared to the IMR vs GT. The results fully reflect the primary evaluation. Results are shown in the following table:

Table 38: 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)			

The results support the primary evaluation.

The Applicant concludes that the accuracy analysis demonstrated that the AI-assisted scores were superior to the manual pathologist scoring for lobular inflammation and hepatocellular ballooning and non-inferior for steatosis and fibrosis. Secondary and exploratory analysis also demonstrated improved accuracy of the AI-assisted scoring compared to manual pathologist evaluation.

This overall conclusion is endorsed, but some questions on clinical meaningfulness on comparison to literature reports remain.

Case studies:

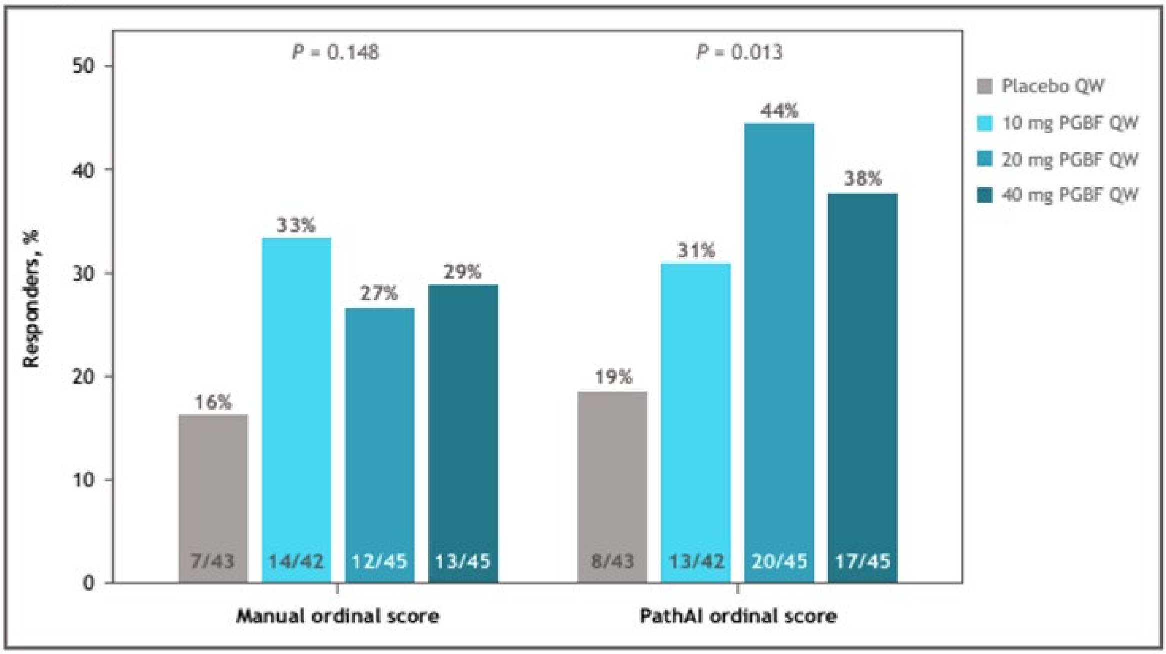
In addition to the actual "validation data", the Applicant has searched for the practical consequences in case the AIM-NASH algorithm is applied to existing data post-hoc, and initially presented 4 studies in which this was explored. These studies comprise trials that were also used for the validation exercise (such as the FALCON 1 study, which was the phase 2b study for Pegbelfermin in stage 3 liver fibrosis, one phase 2 study of semaglutide in the non-cirrhotic population as well as one in the cirrhotic population, but also of the phase 2 study with resmetirom).It is important to note that, in these case studies, the results reflect use of the AIM-NASH algorithm alone, with no pathologist review.

A re-evaluation of study results with application of the AIM-NASH tool has been performed and the results are given in the following per study:

- FALCON 1 study:

Primary endpoint was responders defined as patients with ≥1-stage improvement in NASH CRN fibrosis stage without NASH worsening, or NASH improvement with no worsening of fibrosis, at Week 24.

Figure 8: 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, non-alcoholic steatohepatitis; PGBF, pegbelfermin; QW, once weekly



As shown in the graph, statistically significant differences were only detected in the AIM-NASH based evaluation, but not with the manual evaluation, although effects of the compound were also visible in the manual evaluation.

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

- NCT02912260 (Resmetirom phase 2 study; Madrigal Pharmaceuticals)

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. Results are shown in the following table:

Table 39: AIM-NASH vs. Pathologist detection of endpoint response in Ph2 study of MGL-3196 for

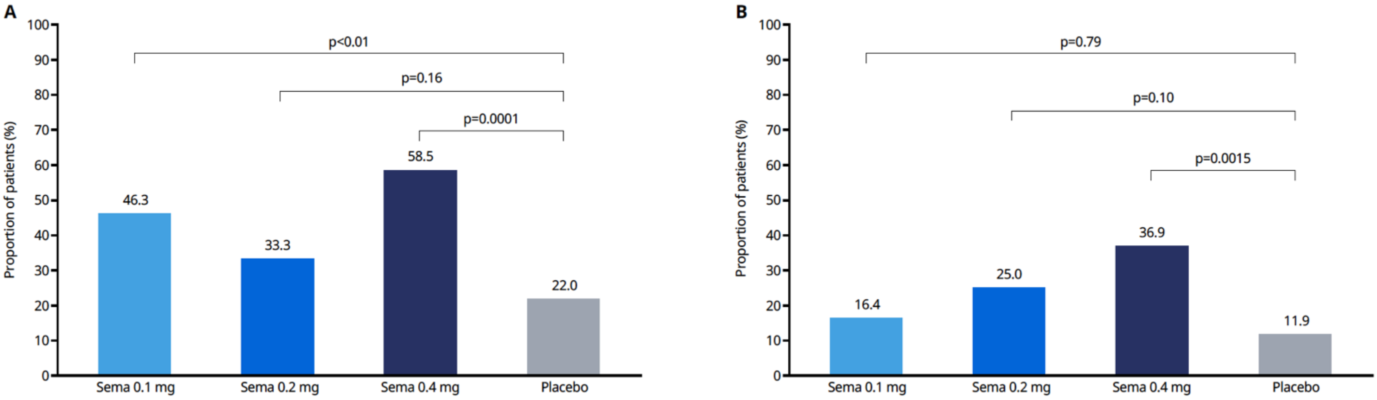
treatment of NASH with CRN Fibrosis Stages 1-3. Consistent with the Central Pathologist and Reader 2, AIM-NASH detected a significantly greater treatment response in the resmetirom-treated group relative to placebo.

In this case, the overall results did not change in any aspect since all, the AIM-NASH, as well as the two manual reader evaluations detected statistically significant differences.

- NCT02970942 (Novo Nordisk phase 2 study for semaglutide with non-cirrhotic population (fibrosis stage 1-3))

The endpoint evaluated was the proportion of patients with NASH resolution without fibrosis worsening. The results of the evaluations are shown in the following figure:

Figure 9: 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. Endpoint: NASH resolution without worsening of fibrosis.

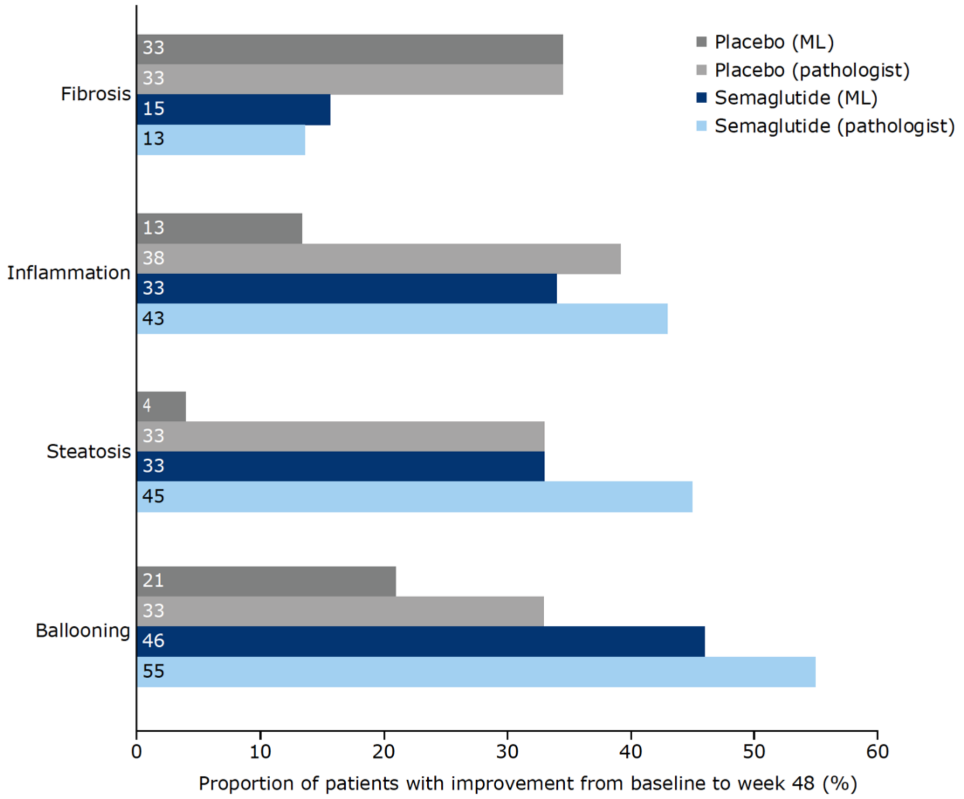


AIM-NASH detected a dose-related response in subjects treated with semaglutide, where increasing dosages of semaglutide resulted in increasingly improved drug response. Of interest seems the fact that results did not only "improve" (dose-response detected), but also "deteriorate", since the effect of the lower doses of semaglutide were relevantly smaller with the AIM-NASH use.

- Study NCT03987451 (Novo Nordisk; Semaglutide phase 2 study in the cirrhotic population)

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.

Figure 10 : 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



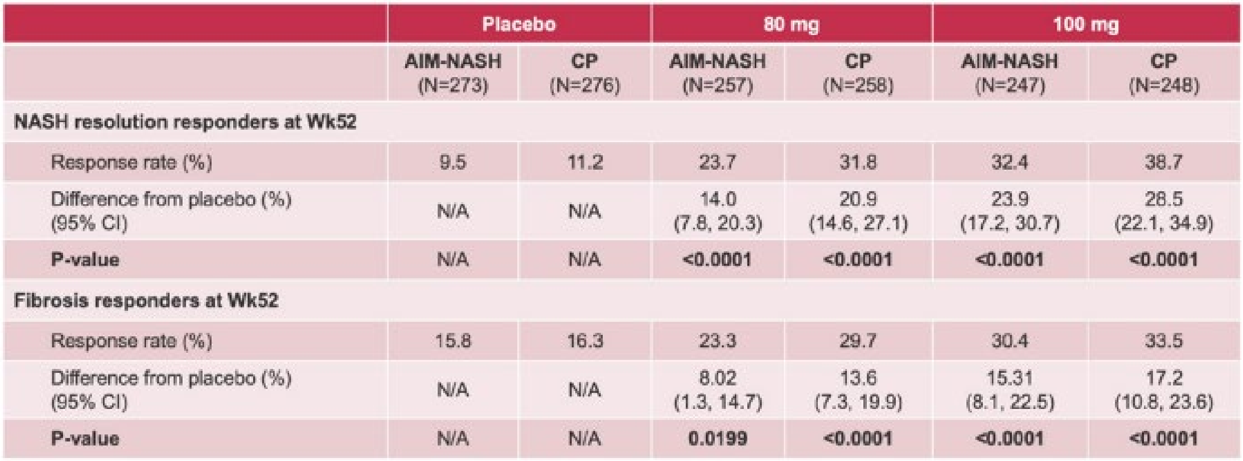
Similar to the other semaglutide study, there is no uniform effect of applying the AIM-NASH tool to the results, since the (negative) conclusion on the effect on fibrosis did not change, and the high response rates in the pathologist ratings for inflammation, steatosis and ballooning were all reduced (however, success rates in placebo were also greatly reduced, so that overall conclusions remained the same).

During the evaluation of the dossier, and as a response to the first List of Issues submitted, the Applicant provided one further "case study". This study (the Resmetirom phase 3 trial) was referenced with a poster presentation at the American Association for the Study of Liver Diseases Meeting in 2023 (Janani S. Iyer, Artificial Intelligence-based Measurement of NASH Histology (AIM-NASH)) and recapitulates primary results from Phase 3 study of resmetirom for treatment of NASH/MASH with liver fibrosis. The main trial results have meanwhile also been published in the New England Journal of Medicine (Harrison SA et al 2024); however, the AIM-NASH evaluations are not included in this publication.

The trial included patients with NAS ≥4 (with at least one in each feature), and fibrosis stages 1b, 2, and 3). Of the total of 966 patients included, 782 and 777 biopsies were evaluated by the "manual" evaluation as well as with the AIM-NASH tool at baseline and at week 52 interim analysis, respectively. Cochran-Mantel-Haenszel (CMH) test stratified for Type 2 diabetes status and baseline fibrosis stage was used to assess statistical significance. Patients were randomized 1:1:1 to resmetirom 80mg, resmetirom 100mg, or placebo administered once daily. Dual primary endpoints at Week 52 were 1) achievement of NASH resolution with no worsening of fibrosis, or 2) ≥1-stage improvement in fibrosis with no worsening of NAFLD Activity Score (NAS).

In the trial, as reported previously (e.g. Harrison 2024) both endpoints were met with high statistical significance. As shown in the following figure, both central pathologist evaluations, as well as the AIM-NASH evaluation showed clear effects on histologic responses.

Figure 11 *:* Treatment vs. placebo response rates measured by the study's central pathologists (CO) vs. AIM NASH per endpoint in a phase 3 study of resmetirom for treatment of NASH/MASH.



	Placebo		80 mg		100 mg	
	AIM-NASH (N=273)	CP (N=276)	AIM-NASH (N=257)	CP (N=258)	AIM-NASH (N=247)	CP (N=248)
NASH resolution responders at Wk52						
Response rate (%)	9.5	11.2	23.7	31.8	32.4	38.7
Difference from placebo (%) (95% CI)	N/A	N/A	14.0 (7.8, 20.3)	20.9 (14.6, 27.1)	23.9 (17.2, 30.7)	28.5 (22.1, 34.9)
P-value	N/A	N/A	**<0.0001**	**<0.0001**	**<0.0001**	**<0.0001**
Fibrosis responders at Wk52						
Response rate (%)	15.8	16.3	23.3	29.7	30.4	33.5
Difference from placebo (%) (95% CI)	N/A	N/A	8.02 (1.3, 14.7)	13.6 (7.3, 19.9)	15.31 (8.1, 22.5)	17.2 (10.8, 23.6)
P-value	N/A	N/A	**0.0199**	**<0.0001**	**<0.0001**	**<0.0001**

For this presentation, it has to be noted that the evaluation is somewhat at odds with the evaluation presented in the NEJM publication, which reported the evaluation of the obvious mITT population (n=318, 316, and 321 in the three treatment groups, summing up to 954 patients, which is "closer" to the "true" ITT population of 966 patients). It is unclear whether the evaluation presented at the 2023 Liver meeting was based on a selection of slides, and on which criteria this selection was based. Nevertheless, the primary interest in this evaluation is the observed differences between the two methods of evaluation, which again demonstrates somewhat lower response rates both in the active, as well as the placebo groups. Obviously, the differences between the active groups become a little bit clearer, but results are overall in full agreement with the primary evaluation.

The "case studies" are considered a valuable contribution to the overall evidence presented and are considered reassuring. It can be shown that results of trials are partly relevantly changed (e.g. better dose-response curves as shown in Figure 8 and 9), but it also appears that the changes occur on both directions, and there is therefore no suspicion of introducing systematic bias into the study results when the AIM-NASH tool is used.

Life-cycle management

The Applicant was asked to describe the life-cycle management in full detail, as it was not clear how the performance of the tool will be monitored after qualification, and how any changes to the tool will be handled.

The Applicant noted that as part of lifecycle management, PathAI has an SOP describing their Change Management Procedure. A major change is any change that affects safety and/or effectiveness. Examples of such changes may include, but are not limited to:

- Change in Context of Use
- Launch in new country
- New user requirements
- New product features
- Enhancements to existing features
- Change in QC methods

- Security changes
- Equipment changes

This SOP covers the retesting requirements for the situations as reflected in the above list and aligns with requirements outlined in ISO 13485 as well as relevant sections of FDA regulations (21 CFR 820).

When a change is proposed, consideration is given to the potential impact of the change on function, performance, usability, safety and risk, and applicable regulatory requirements. Planning for changes includes evaluation of both direct and indirect impacts that the change could have. Justification for the changes and how they will be implemented is documented via a software change request form.

This approach is in principle supported for the time being. However, one concern was that small changes might also accumulate over time and require some revalidation. The Applicant stated that the involvement of a pathologist in the AIM-NASH workflow will help in the evaluation of such changes to ensure the continued safe and effective use of the product. Per PathAI procedure, regulatory assessment of proposed changes is required and includes evaluation of the impact of current proposed modifications considered together with any previous modifications since last submitted to a regulatory body to ensure any accumulation has not resulted in significant changes or modifications that require premarket notification.

According to the Applicant’s proposal, any changes that impact the AIM-NASH tool’s safety and effectiveness will be evaluated and the necessary verification and/or validation will be documented and are intended to be submitted to EMA before release. In addition to evaluating intentional updates and changes made to the tool, PathAI will also monitor performance of the tool as deployed in clinical trials. Agreements with biopharma partners will include language allowing PathAI to survey AIM-NASH performance once deployed in trials. Tool performance will be compared to the validated tool and significant deviation from validated performance will be investigated and understood by PathAI. Performance changes that could impact the safety and effectiveness of AIM-NASH are intended to be reported to EMA via email initially, then continue as appropriate based on discussions. This approach is, in principle, supported.

However, since the current qualification relates to the tool at the “data lock point”, and a variation of qualification opinions is not foreseen, all changes that occurred between the current submission date, and the time of when the tool is used during a development programme for a distinct medicinal product will need to be submitted within the MAA of that medicinal product by the respective cooperation partner of PathAI (the intended MA holder). This will need to include the documentation and assessment of impact of all changes to the tool, both minor, “accumulated minor” as well as major changes.

However, in case several “major” (or other accumulated) changes arise over time, the submission of the meanwhile generated additional data may be impractical for the evaluation of a specific MAA, and the Applicant is therefore recommended to discuss with the CHMP whether a “re-qualification” of the tool could be the better way forward.

While therefore, on a preliminary basis, the proposed measures to assess changes, according to the submitted documents are acknowledged, a periodical need to re-evaluate the need for “re-validation” and/or “re-qualification” is to be considered.

In addition, there is a principal property of machine learning (ML) models, which appears not to be addressed. These models are at risk of temporal model degradation, where the performance of the model drops over time (see: Reflection paper on the use of Artificial Intelligence (AI) in the medicinal product lifecycle. EMA/CHMP/CVMP/83833/2023). The pace at which the performance drops is usually unknown prior to deployment in real-world settings. Model performance is influenced by several issues including the evolution of the underlying data distributions, and differences in operating system, scanner, procedures, among others. This is a known issue, typically managed by the systems developer through a set of operations called machine learning operations, aimed at structuring workflows, monitoring, and improving the performance of the model which includes re-training of the model if performance falls below a pre-defined threshold. It is advisable to have such a framework in place for this purpose in the future.

There are considerations related to continuous monitoring model performance, and especially AI model degradation, that are potentially not covered by the proposed SOP and/or the details of the technical ISO standard and FDA regulations referred to and which will need to be considered.

The complexity of AI model degradation and change management, however, is a significant scientific challenge that has not yet been satisfactorily resolved from a (quantitative) methodology perspective. Technological and methodological advancements present significant opportunities, but these complex models, despite their high performance, may not always generalise well in the future. There are uncertainties regarding the validity and performance of these tools in the future – including learning and future data use elements – necessitating further monitoring and evaluation.

In this context predetermined change control plans to be developed have been put on the table for discussion in other jurisdictions, but overall, no specific guidance is available yet that can serve as an established reference framework for the life-cycle management of AI based tools.

Consequently, for the time being, and as repeatedly requested in the report, a pragmatic alternative is to declare that both, continuous model performance and future changes to the methodology, which is qualified in this procedure, will need to be well documented and reported within the dossier of the medicinal product for which it is used. In case of major changes, re-qualification might be considered as a more efficient way forward.

In the future, the Applicant is strongly encouraged contribute to delivering the science and rationale regarding life-cycle management of AI based tools. Based on this, the Applicant could develop a full scientific justification on why their SOP for change management is adequate and robust, which would foster dialogue and scientific discussion.

Overall conclusion:

The proposed AIM-NASH tool has been adequately developed and evaluated for its performance within the defined Context of Use in comparison to manual reading of histology and has demonstrated at least similar variability, and partly improved variability in certain features of NASH histology assessment. Overall, qualification of the tool is recommended.

Appendices to be published:

1. User Manual Version 1.5, dated 2023-11
2. Briefing document
3. Written responses to first List of Issues
4. Written responses to second List of Issues
5. Overview of comments received on Draft Qualification Opinion

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).

37. Brunt, Elizabeth M., Andrew D. Clouston, Zachary Goodman, Cynthia Guy, David E. Kleiner, Carolin Lackner, Dina G. Tiniakos, et al. 2022. 'Complexity of Ballooned Hepatocyte Feature Recognition: Defining a Training Atlas for Artificial Intelligence-Based Imaging in NAFLD'. *Journal of Hepatology* 76 (5). https://doi.org/10.1016/j.jhep.2022.01.011.

38. Cicchetti, Domenic V., and Truett Allison. 1971. 'A New Procedure for Assessing Reliability of Scoring EEG Sleep Recordings'. *American Journal of EEG Technology* 11 (3). https://doi.org/10.1080/00029238.1971.11080840.

39. Davison, Beth A., Stephen A. Harrison, Gad Cotter, Naim Alkhouri, Arun Sanyal, Christopher Edwards, Jerry R. Colca, Julie Iwashita, Gary G. Koch, and Howard C. Dittrich. 2020. 'Suboptimal Reliability of Liver Biopsy Evaluation Has Implications for Randomized Clinical Trials'. *Journal of Hepatology* 73 (6). https://doi.org/10.1016/j.jhep.2020.06.025.

40. Iyer, Janani, Pierre Bedossa, Cynthia Guy, Hang Zhang, Brian Baker, Darren Fahy, Tayla Parker-Shen, et al. 2023. 'Artificial Intelligence-Based Measurement of NASH Histology (AIM-NASH) Recapitulates Primary Results from Phase 3 Study of Resmetirom for Treatment of NASH/MASH with Liver Fibrosis'. In *American Association for the Study of Liver Diseases*. Boston, MA.

41. Kleiner, David E., Elizabeth M. Brunt, Mark Van Natta, Cynthia Behling, Melissa J. Contos, Oscar W. Cummings, Linda D. Ferrell, et al. 2005. 'Design and Validation of a Histological Scoring System for Nonalcoholic Fatty Liver Disease'. *Hepatology* 41 (6). https://doi.org/10.1002/hep.20701.

42. Kleiner, David E., Elizabeth M. Brunt, Laura A. Wilson, Cynthia Behling, Cynthia Guy, Melissa Contos, Oscar Cummings, et al. 2019. 'Association of Histologic Disease Activity With Progression of Nonalcoholic Fatty Liver Disease'. *JAMA Network Open* 2 (10). https://doi.org/10.1001/jamanetworkopen.2019.12565.

43. Sanyal, Arun, Rohit Loomba, Quentin Anstee, Vlad Ratziu, Amrik Shah, Macky Natha, Deepa Rajagopalan, et al. 2021. 'Minimizing Variability and Increasing Concordance for NASH Histological Scoring in NASH Clinical Trials'. In *American Association for the Study of Liver Diseases*.

44. Pulaski, Hanna, Shraddha S Mehta, Laryssa C Manigat, Stephanie Kaufma , Hypatia Hou, ILKe Nalbantoglu , Xuchen Zhang, Emily Curl, Ross Taliano , Tae Hun Kim 5, Michael Torbenson 6, Jonathan N Glickman, Murray B Resnick, Neel Patel, Cristin E Taylor, Pierre Bedossa, Michael C Montalto, Andrew H Beck, Katy E Wack. Validation of a whole slide image management system for metabolic-associated steatohepatitis for clinical trials. *Journal of Pathology: Clinical Research. Sept 2021. https://doi.org/10.1002/2056-4538.12395*

45. Iyer JS, Juyal D, Le Q, Shanis Z, Pokkalla H, Pouryahya M, Pedawi A, Stanford-Moore SA, Biddle-Snead C, Carrasco-Zevallos O, Lin M, Egger R, Hoffman S, Elliott H, Leidal K, Myers RP, Chung C, Billin AN, Watkins TR, Patterson SD, Resnick M, Wack K, Glickman J, Burt AD, Loomba R, Sanyal AJ, Glass B, Montalto MC, Taylor-Weiner A, Wapinski I, Beck AH. Nature Medicine. Aug 2024. *https://doi.org/10.1038/s41591-024-03172-7*

46. Pulaski, Hanna, Ph.D., Stephen A. Harrison, M.D.2, Shraddha S. Mehta, Ph.D., Arun J Sanyal, M.D., Marlena C. Vitali, BS, Laryssa C. Manigat, Ph.D., Hypatia Hou, BS, Susan P. Madasu Christudoss, BS, Sara M. Hoffman, BA, Adam Stanford-Moore, MS, Robert Egger, Ph.D.,

Jonathan Glickman, M.D., Ph.D., Murray Resnick, M.D., Ph., Neel Patel, Ph.D., Cristin E. Taylor, PA-C, DPT, Robert P. Myers, M.D., Chuhan Chung, M.D., Scott D. Patterson, Ph.D., Anne-Sophie Sejling, M.D., Ph.D., Anne Minnich, Ph.D., Vipul Baxi, MS, MBA, G. Mani Subramaniam, M.D., Ph.D., Quentin M. Anstee, MB BS, PhD, Rohit Loomba, M.D., M.H.Sci, Vlad Ratziu, M.D., PhD., Michael C Montalto, Ph.D., Nick P Anderson, Ph.D., Andrew H Beck, M.D., Ph.D., Katy E Wack, Ph.D. Clinical validation of an AI-based pathology tool for scoring of metabolic dysfunction-associated steatohepatitis. Nature Medicine. Accepted Sept 16, 2024 for November 2024 publication. *https://doi.org/10.1038/s41591-024-03301-2*